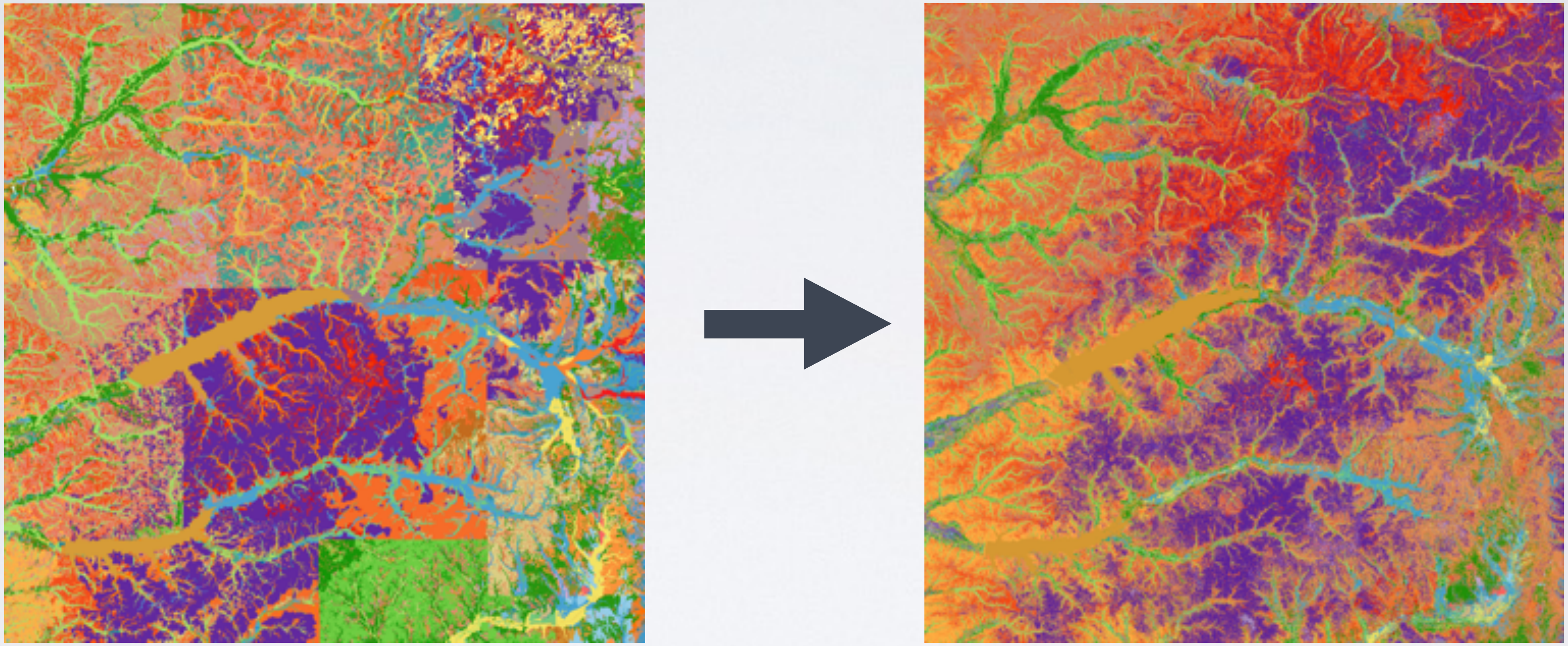
Spatial Disaggregation and Harmonization of gSSURGO



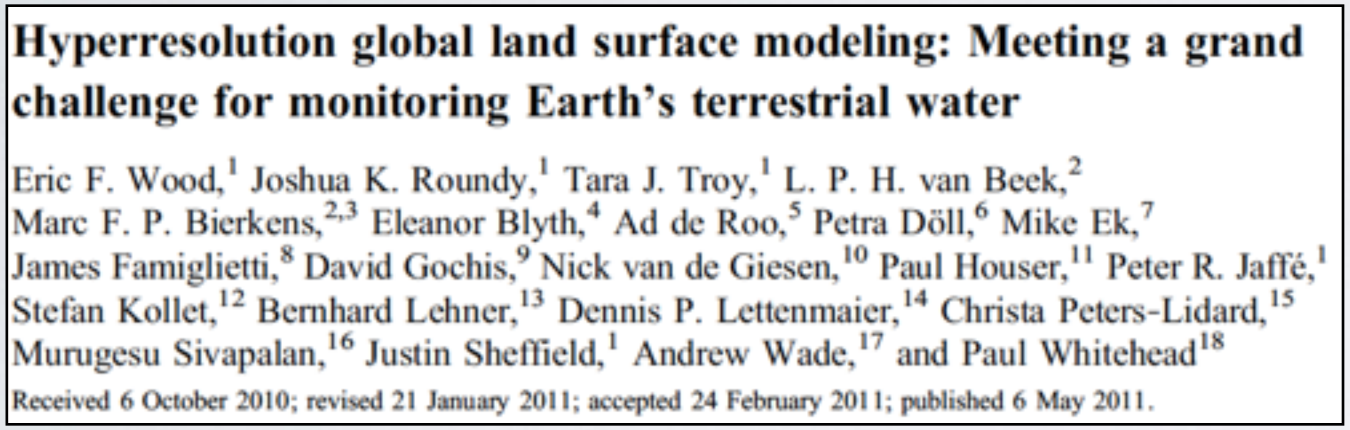
Nathaniel Chaney, Jonathan Hempel,
Nathan Odgers, Alex McBratney, Eric F. Wood

PRINCETON UNIVERSITY

NRCS

THE UNIVERSITY OF SYDNEY

MOTIVATION: NEXT GENERATION LAND SURFACE MODELING



Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water

Eric F. Wood,[1] Joshua K. Roundy,[1] Tara J. Troy,[1] L. P. H. van Beek,[2] Marc F. P. Bierkens,[2,3] Eleanor Blyth,[4] Ad de Roo,[5] Petra Döll,[6] Mike Ek,[7] James Famiglietti,[8] David Gochis,[9] Nick van de Giesen,[10] Paul Houser,[11] Peter R. Jaffé,[1] Stefan Kollet,[12] Bernhard Lehner,[13] Dennis P. Lettenmaier,[14] Christa Peters-Lidard,[15] Murugesu Sivapalan,[16] Justin Sheffield,[1] Andrew Wade,[17] and Paul Whitehead[18]

Received 6 October 2010; revised 21 January 2011; accepted 24 February 2011; published 6 May 2011.

WRR | Water Resources Research

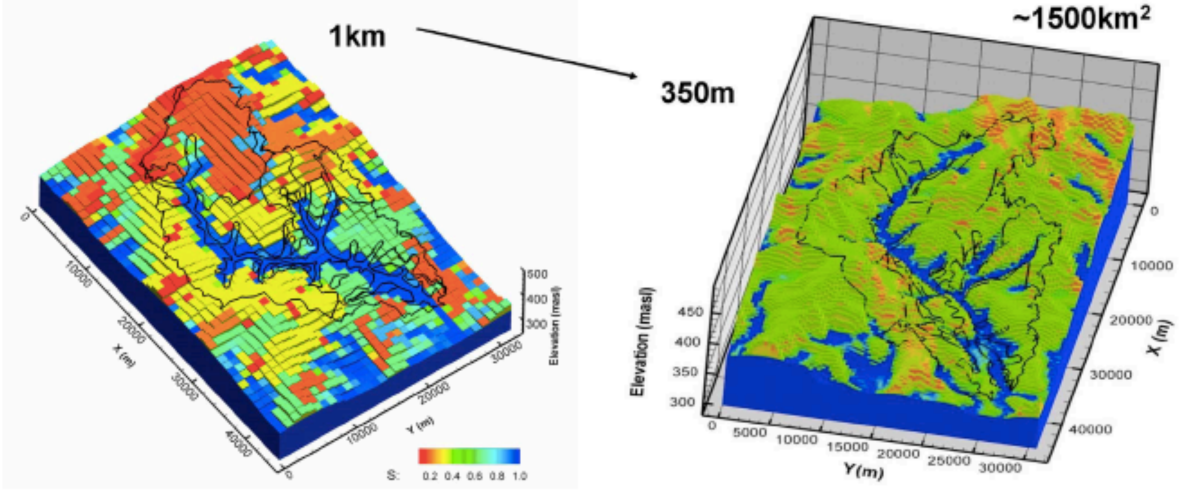


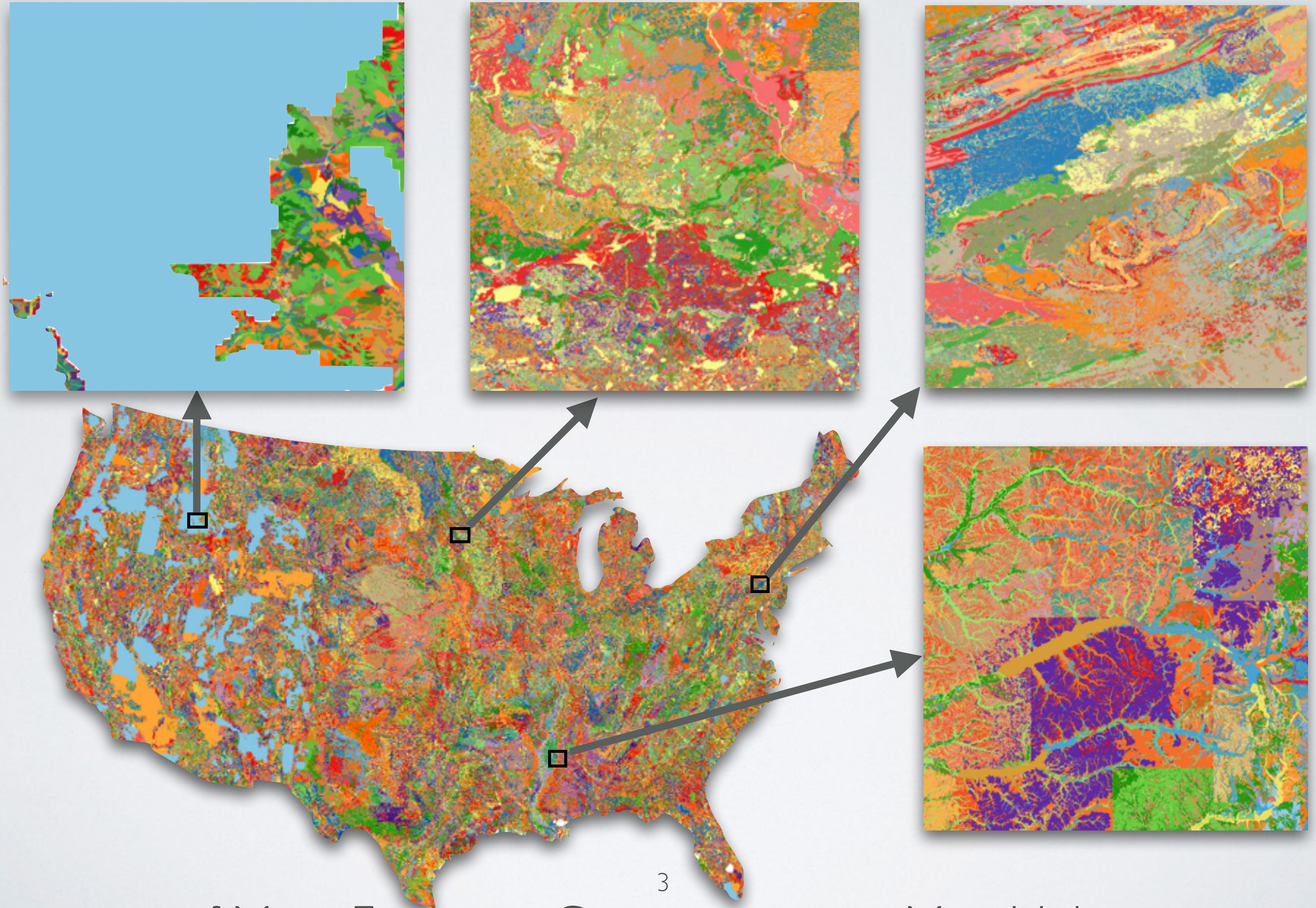
Figure 1. Higher-resolution modeling leads to better spatial representation of saturated and nonsaturated areas, with implications for runoff generation, biogeochemical cycling, and land-atmosphere interactions. Soil moisture simulations on the Little Washita showing the impact that the resolution has on its estimation [*Kollet and Maxwell*, 2008].

Goal: ~100 meters global

Challenges:

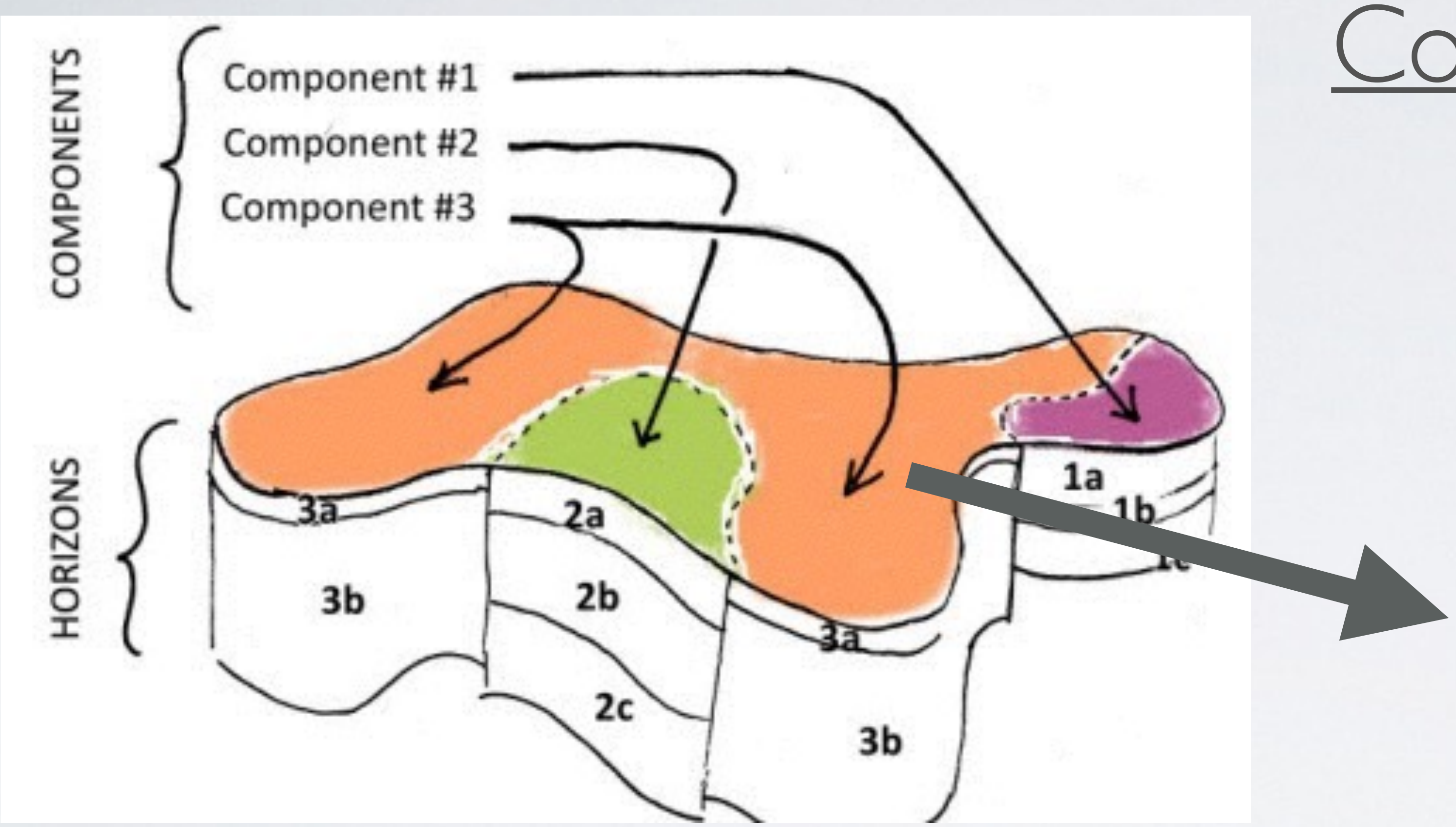
- Model Structure
- Input Data
- Computation

Motivation: gSSURGO Tradeoffs

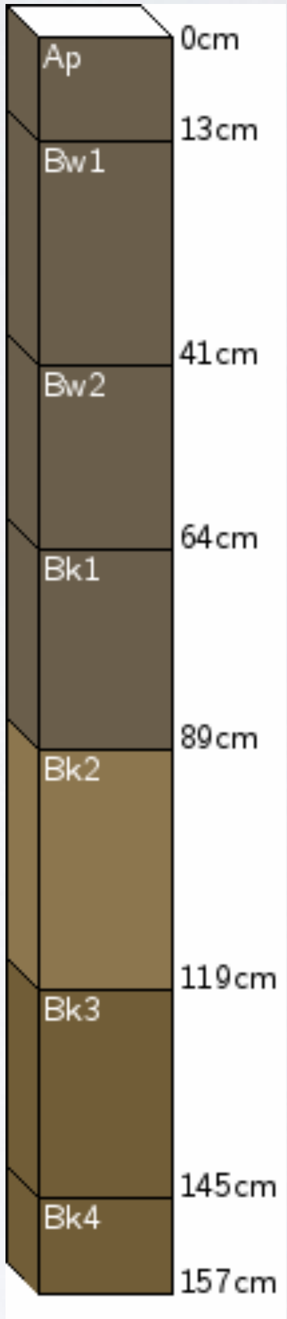


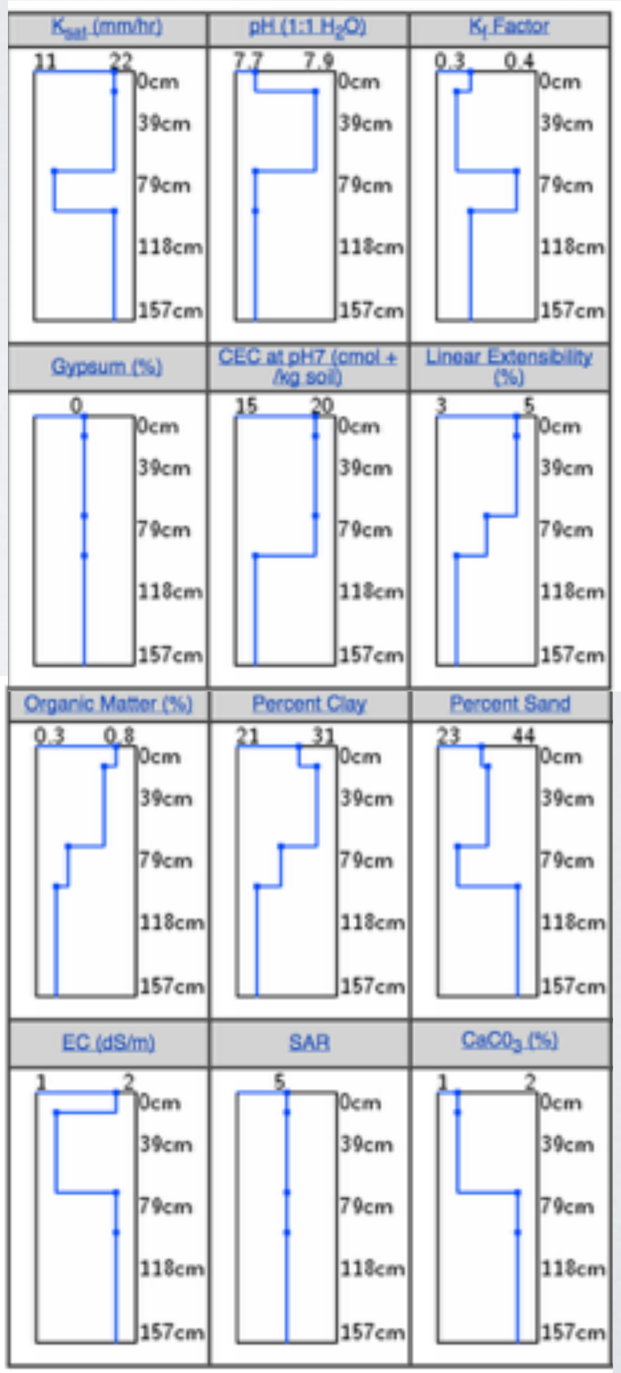
*Most Frequent Component per Map Unit

SSURGO: COMPONENT INFO



Component Name: Cerini

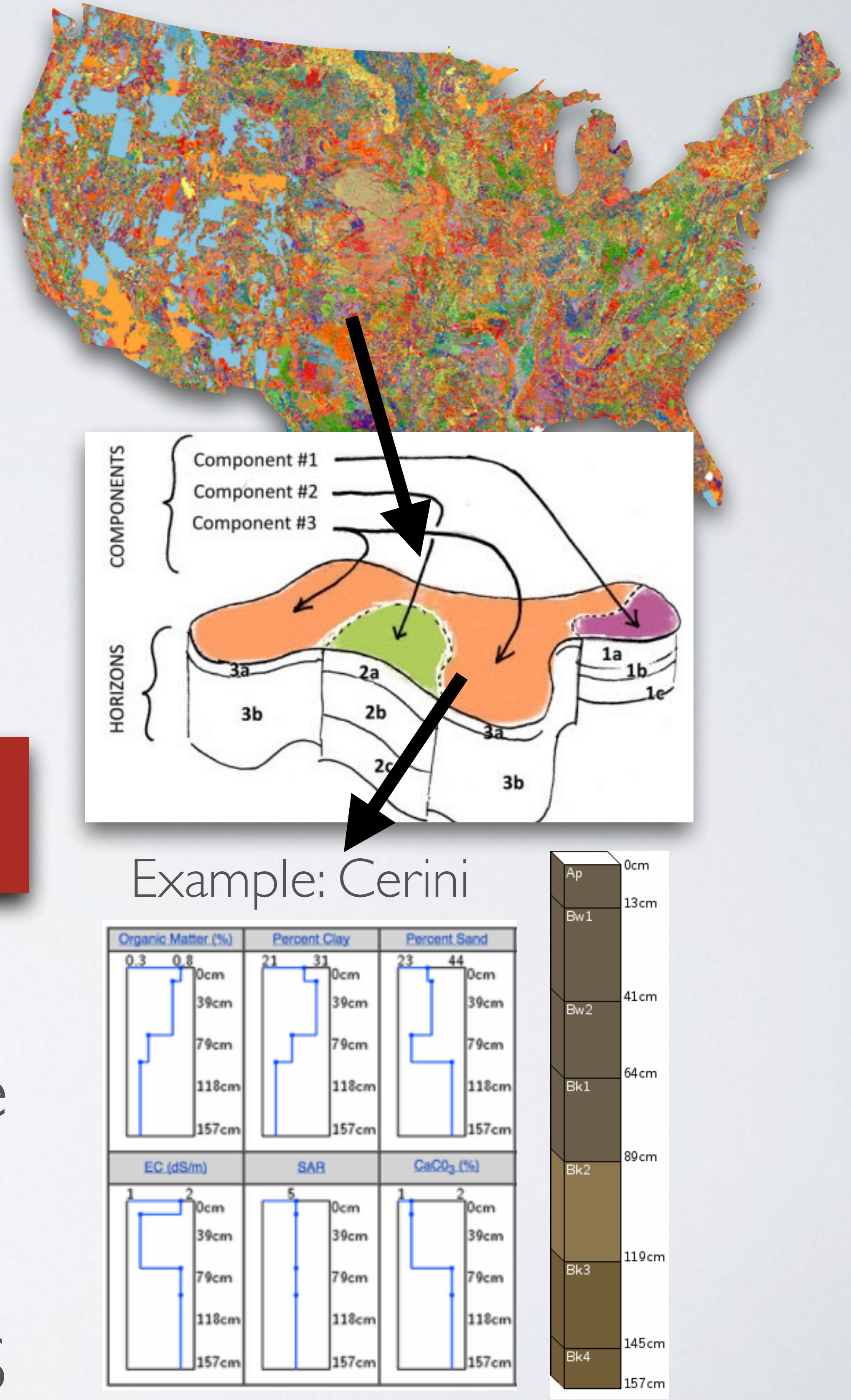




- Rich database per component
- Uncertainty information
  - Triangular Distribution

Motivation and Outline

gSSURGO Tradeoffs	
Advantages	Challenges
Spatial Detail	Survey Bias (Boundaries)
Rich Database	Incomplete
In Situ Observations	Variable Resolution



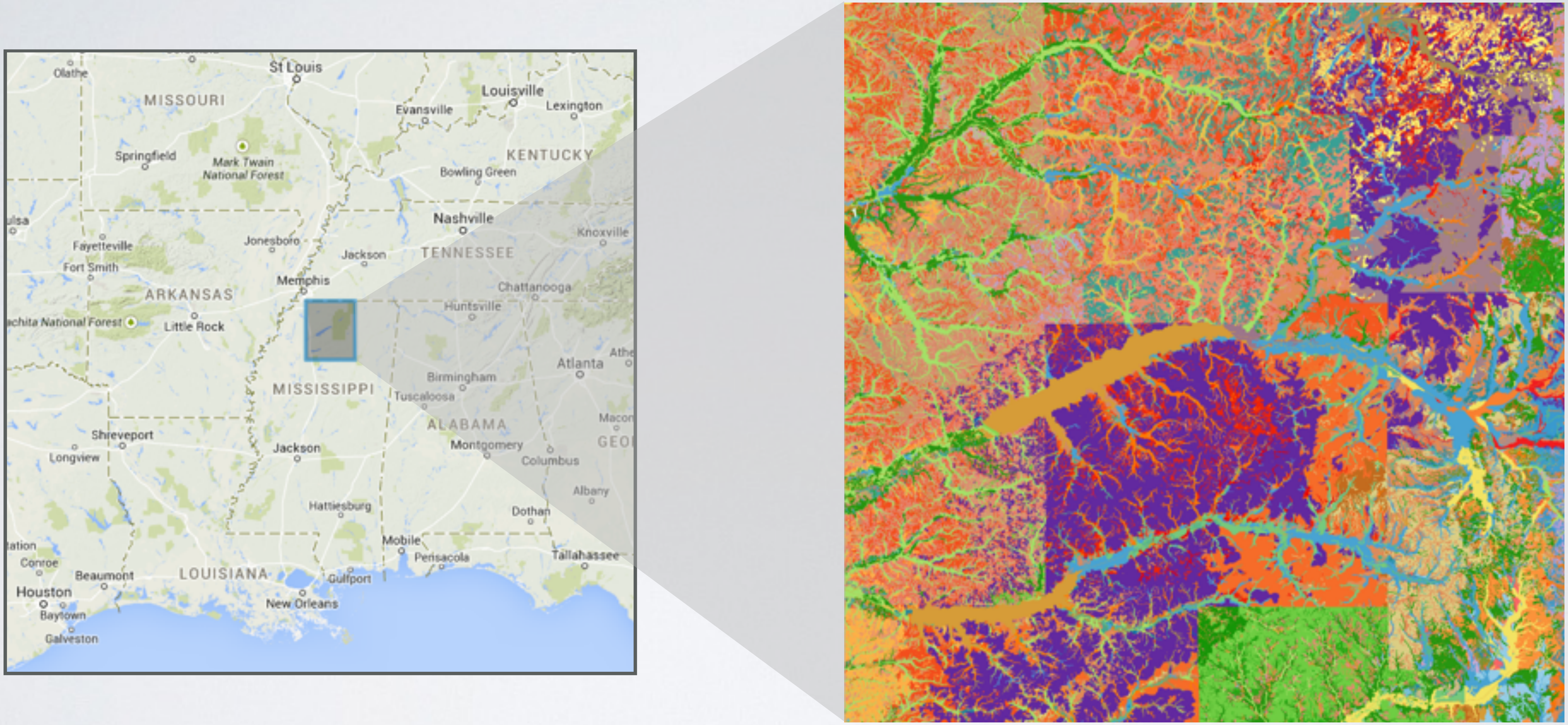
Goal: Address gSSURGO challenges

Outline:

A. Testbed: Northern Mississippi State
B. Methodology: DSMART
C. Application over CONUS (HPC)
D. Explore new dataset over CONUS

Source: http://casoilresource.lawr.ucdavis.edu

Testbed: Northern Mississippi State

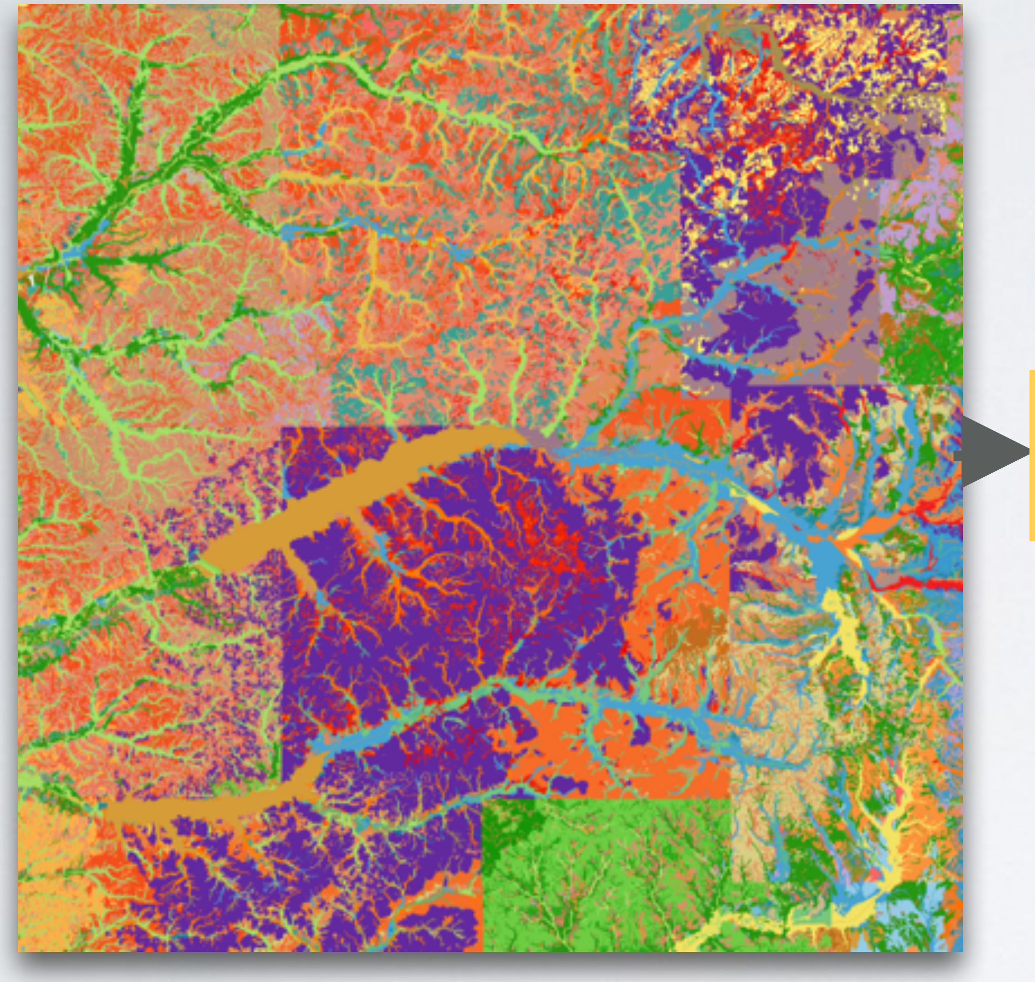


gSSURGO

*Most Frequent Component per Map Unit

Objective

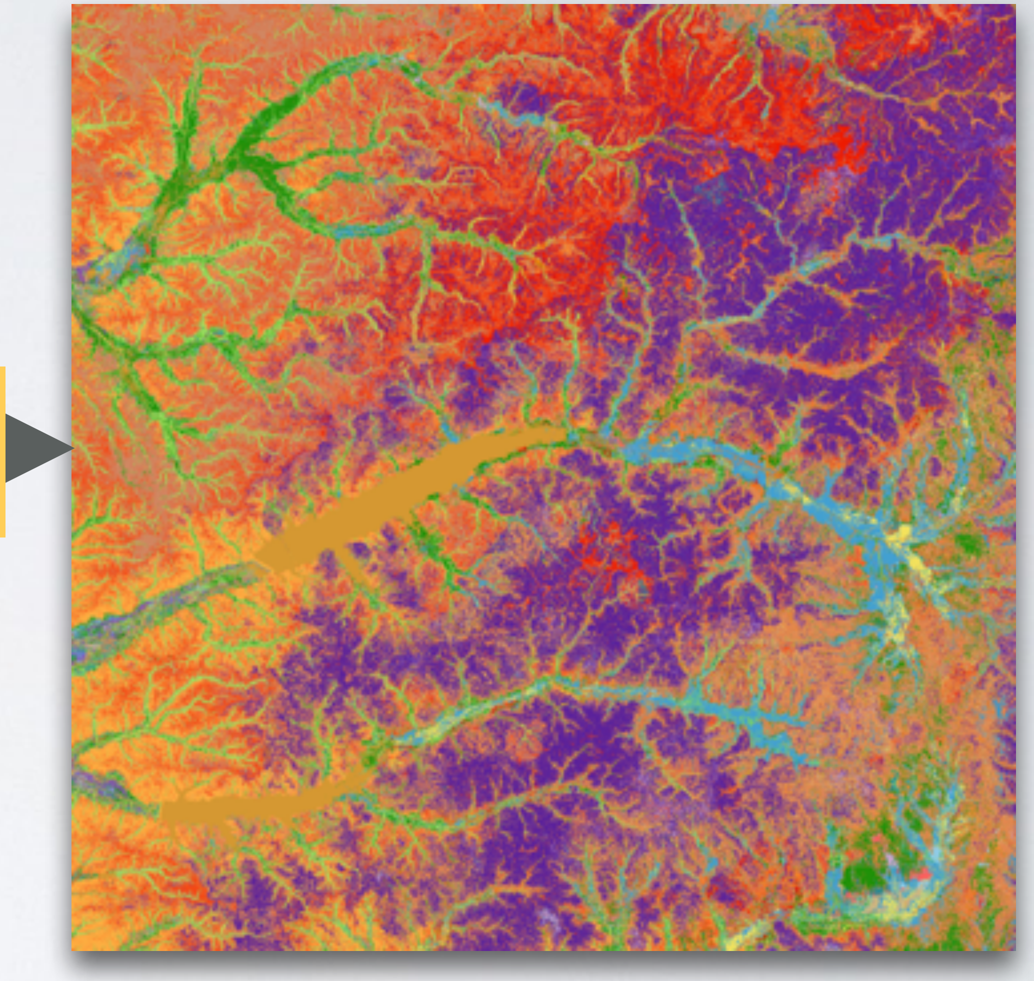
Legacy Soil Data



Algorithm

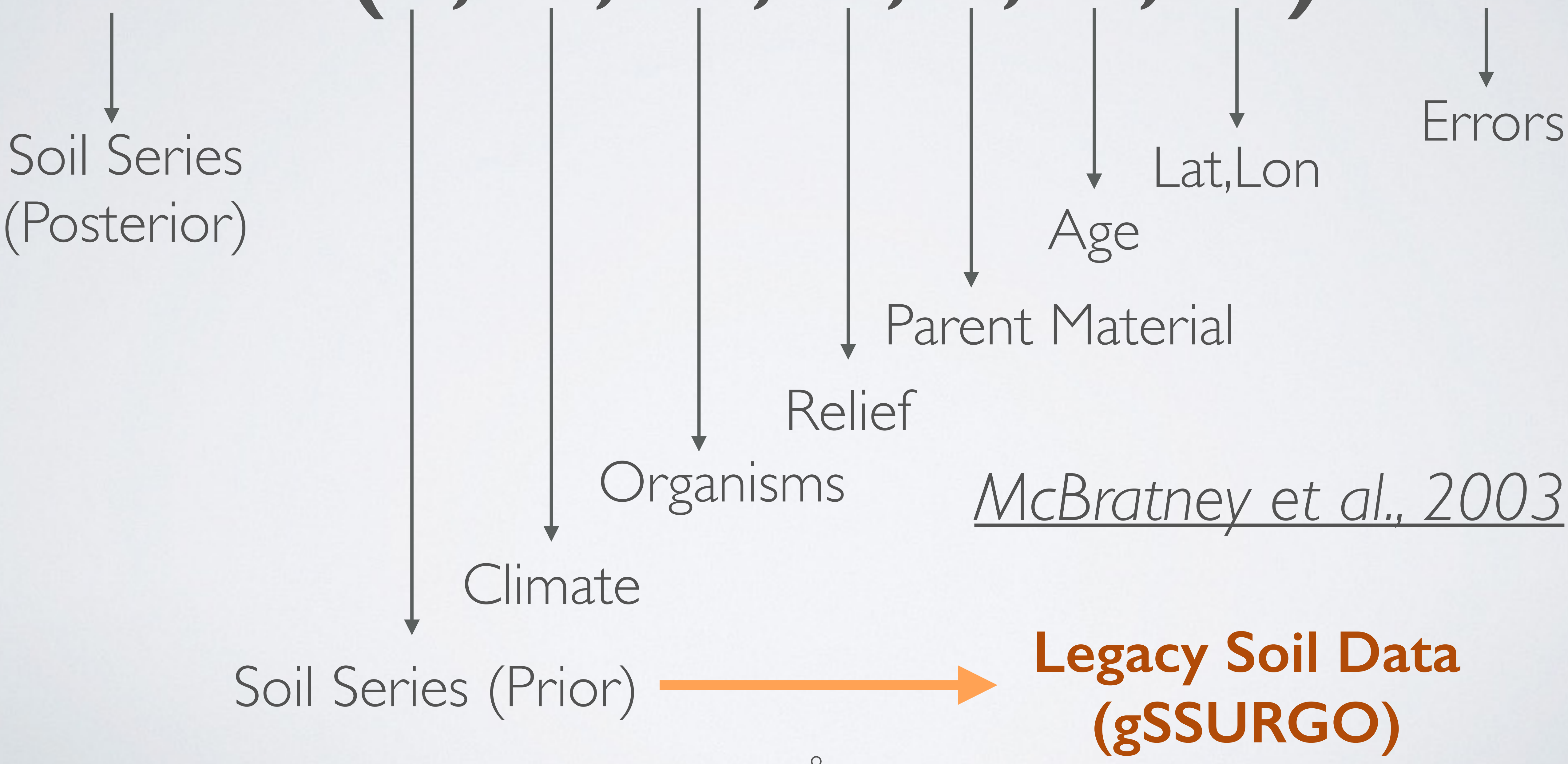
Soil Covariates

Corrected Product



DIGITAL SOIL MAPPING

$$S = f\,(S, C, O, R, P, A, N) + \varepsilon$$



Soil Series (Posterior)

Errors

Lat,Lon

Age

Parent Material

Relief

Organisms

Climate

Soil Series (Prior)

Legacy Soil Data (gSSURGO)

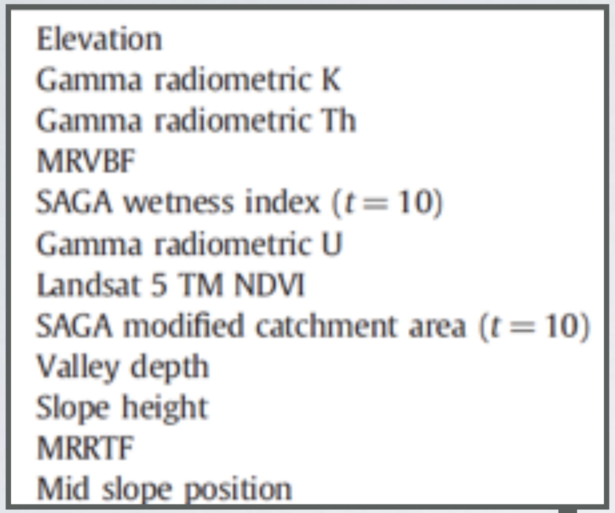
McBratney et al., 2003

Soil Covariates: CONUS

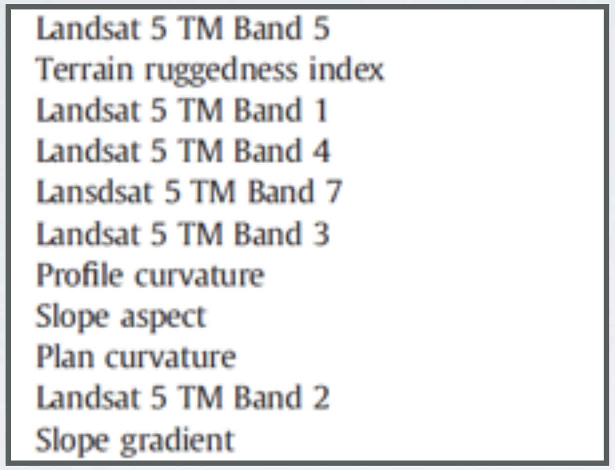
	Dataset	Soil Covariate	Resolution
Relief	NED DEM	Topographic Index Elevation MRVBF MRRTF Curvature Slope Accumulation Area	30 meters
Parent Material	USGS Aeroradiometric	Uranium Thorium Potassium	4000 meters
Organisms	NLCD	Land Cover Type	30 meters

Algorithm: DSMART

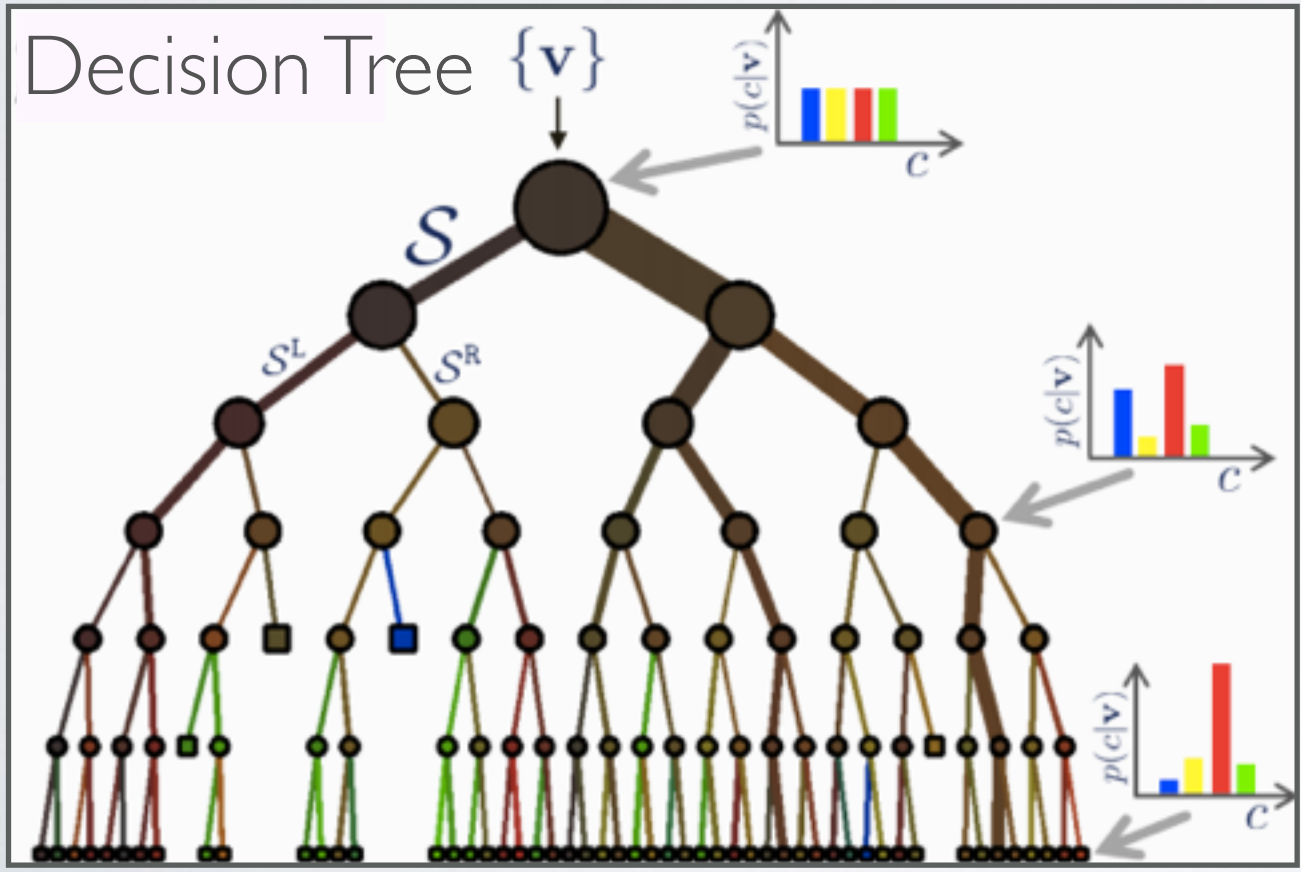
SOIL COVARIATES



Elevation
Gamma radiometric K
Gamma radiometric Th
MRVBF
SAGA wetness index ($t = 10$)
Gamma radiometric U
Landsat 5 TM NDVI
SAGA modified catchment area ($t = 10$)
Valley depth
Slope height
MRRTF
Mid slope position

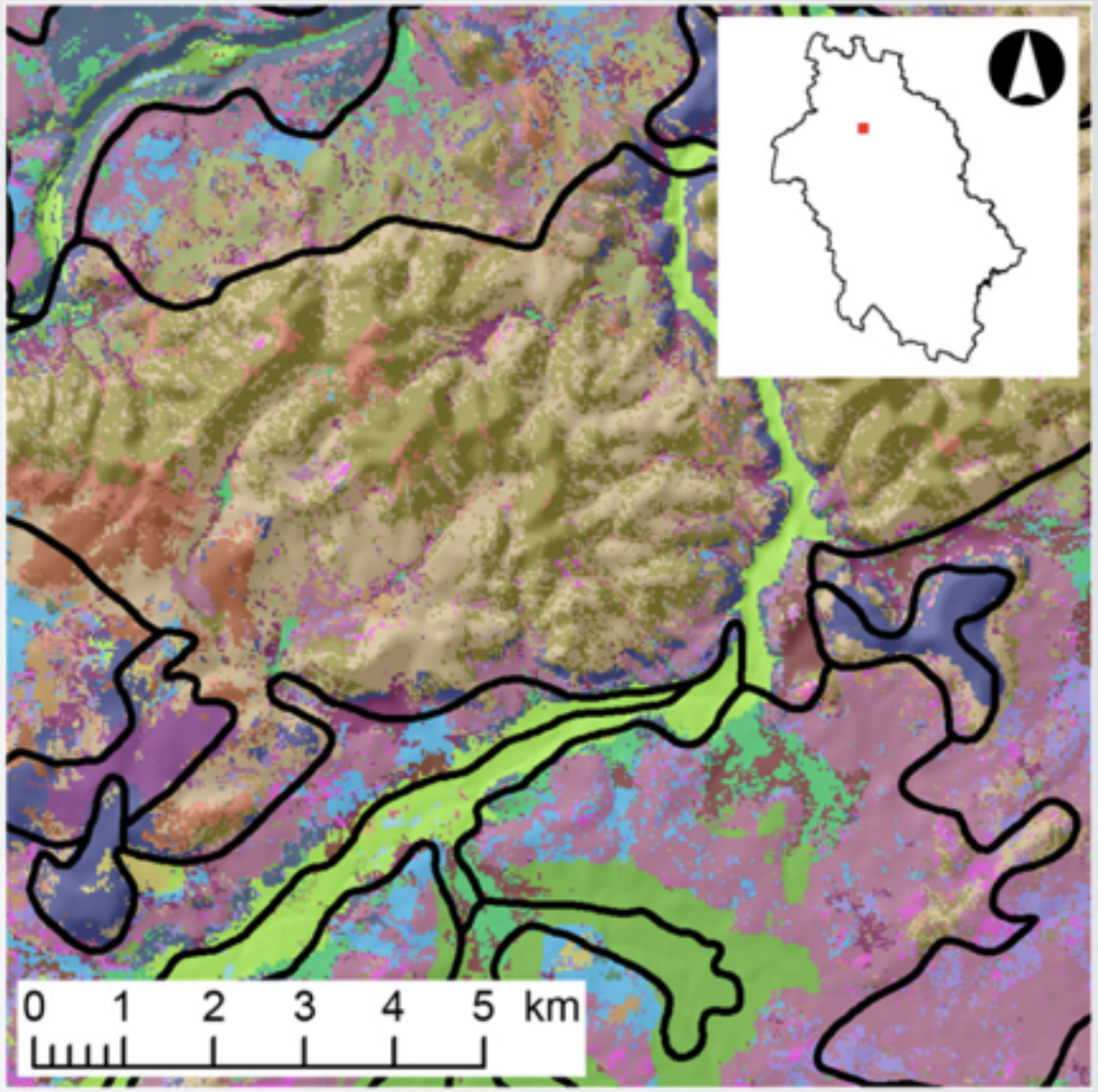


Landsat 5 TM Band 5
Terrain ruggedness index
Landsat 5 TM Band 1
Landsat 5 TM Band 4
Lansdsat 5 TM Band 7
Landsat 5 TM Band 3
Profile curvature
Slope aspect
Plan curvature
Landsat 5 TM Band 2
Slope gradient



Decision Tree

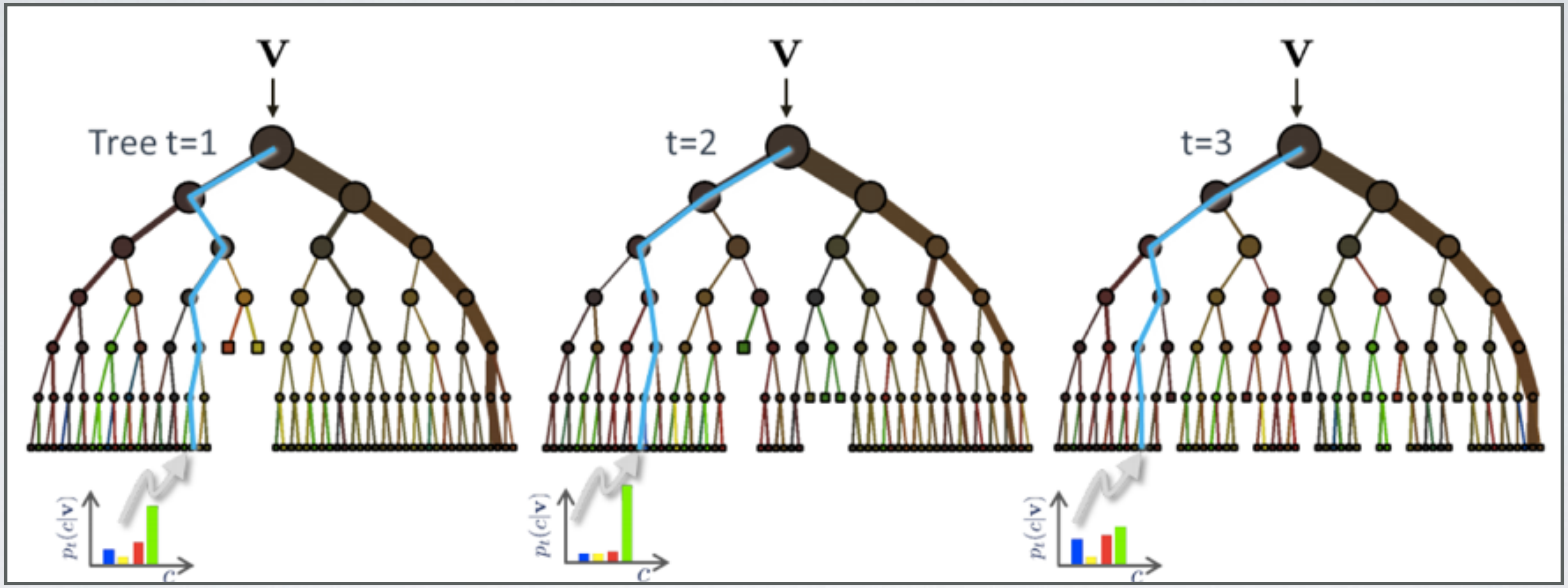
Source: Microsoft Research



Source: Odgers et al., 2014

Train with legacy soil data

Enhanced DSMART: Random Forest

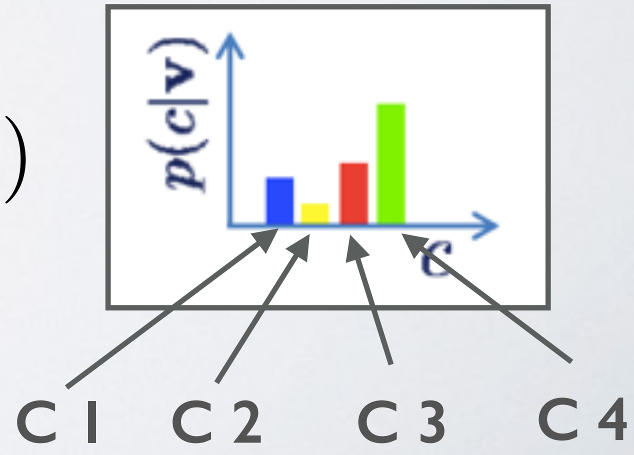


Source: Microsoft Research

Soil Covariates

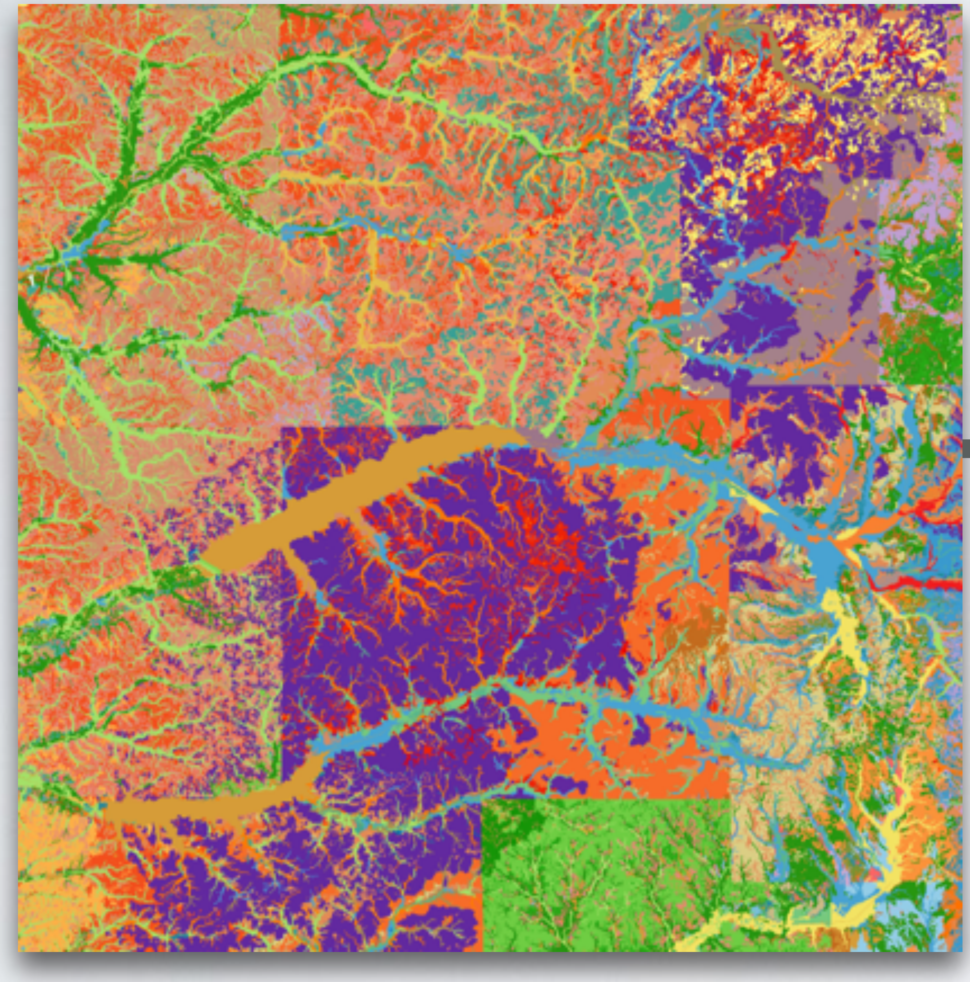
Forest output probability: $p(c|\mathbf{v}) = \frac{1}{T}\sum_{t}^{T} p_t(c|\mathbf{v})$

Component



Enhanced DSMART: Result

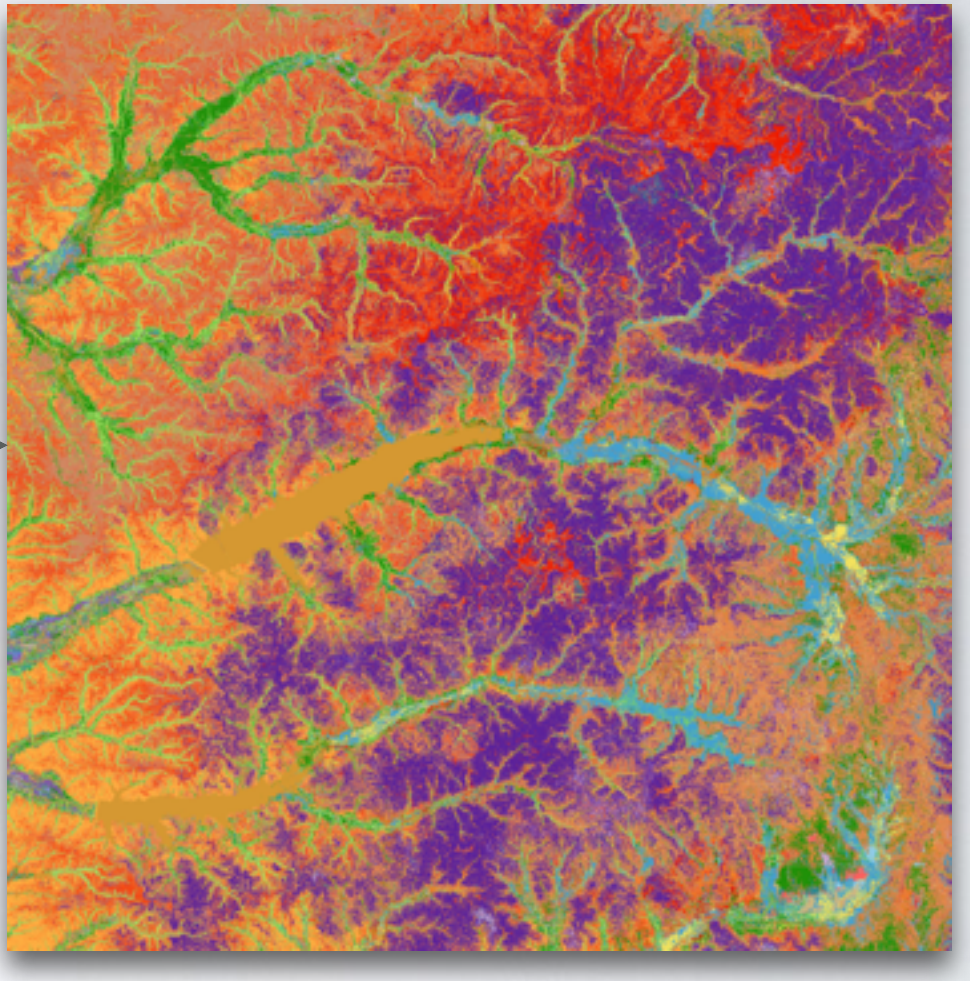
gSSURGO



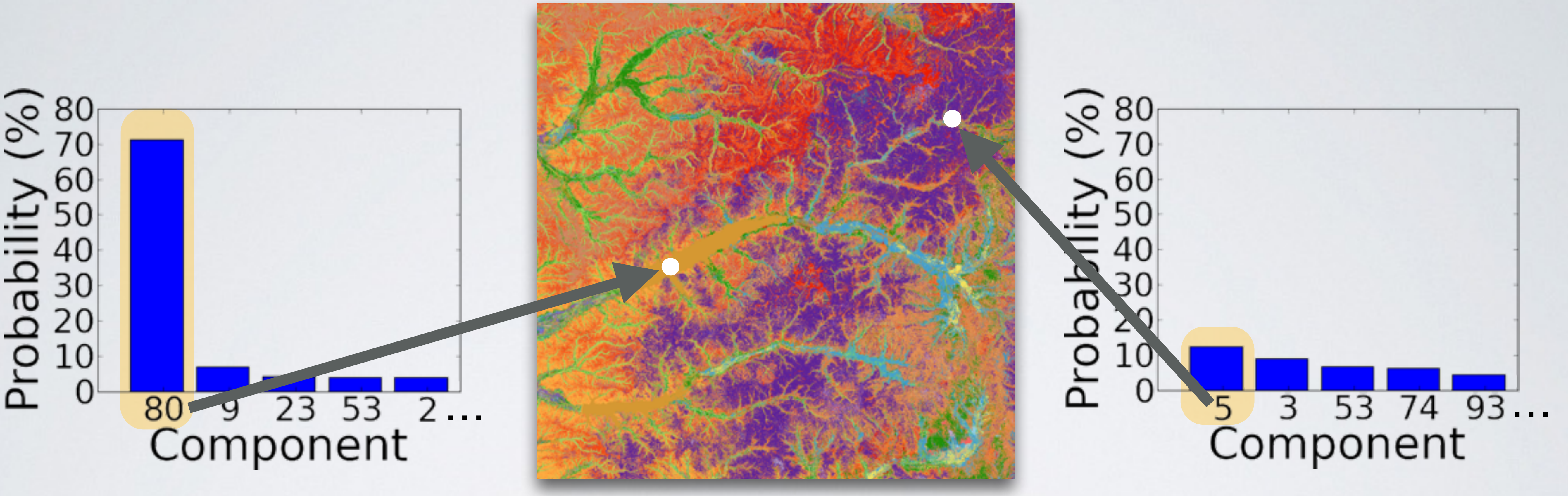
DSMART

Soil Covariates

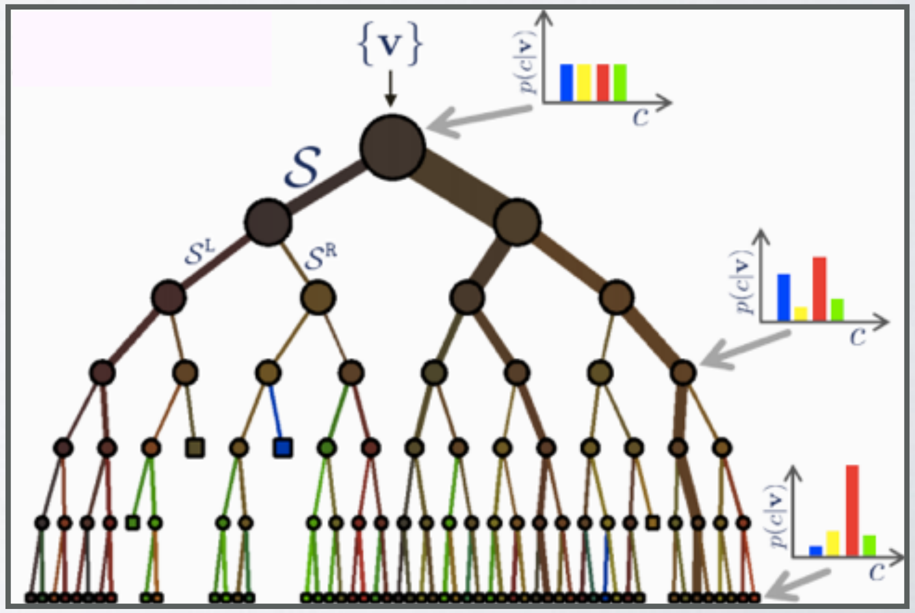
Corrected Product



Detailed Info: Probabilities

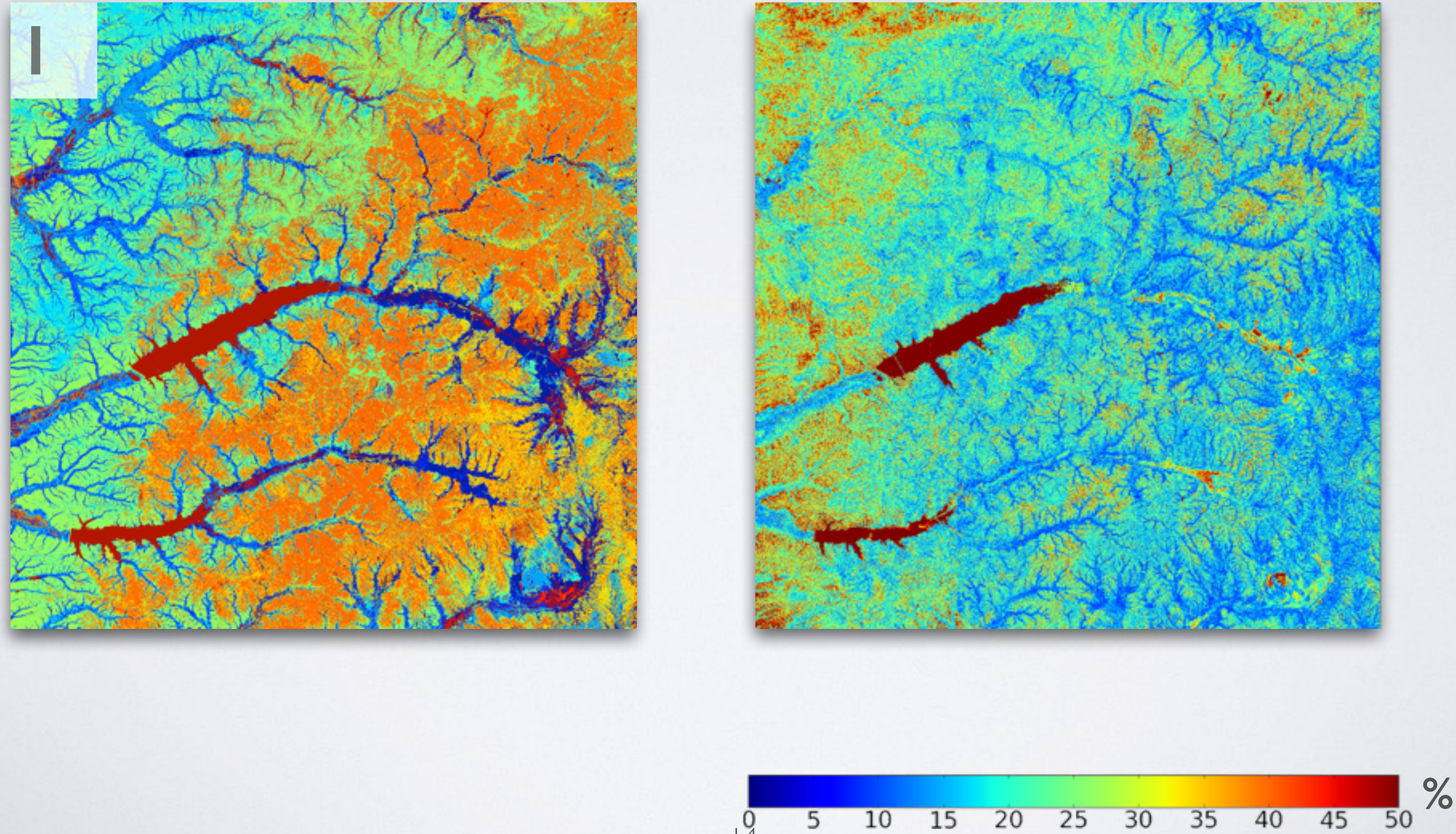


- Decision Tree Leaf - Component Histogram
  A. Each grid cell (soil covariates) falls on a leaf
- Implication → Quantify component uncertainty



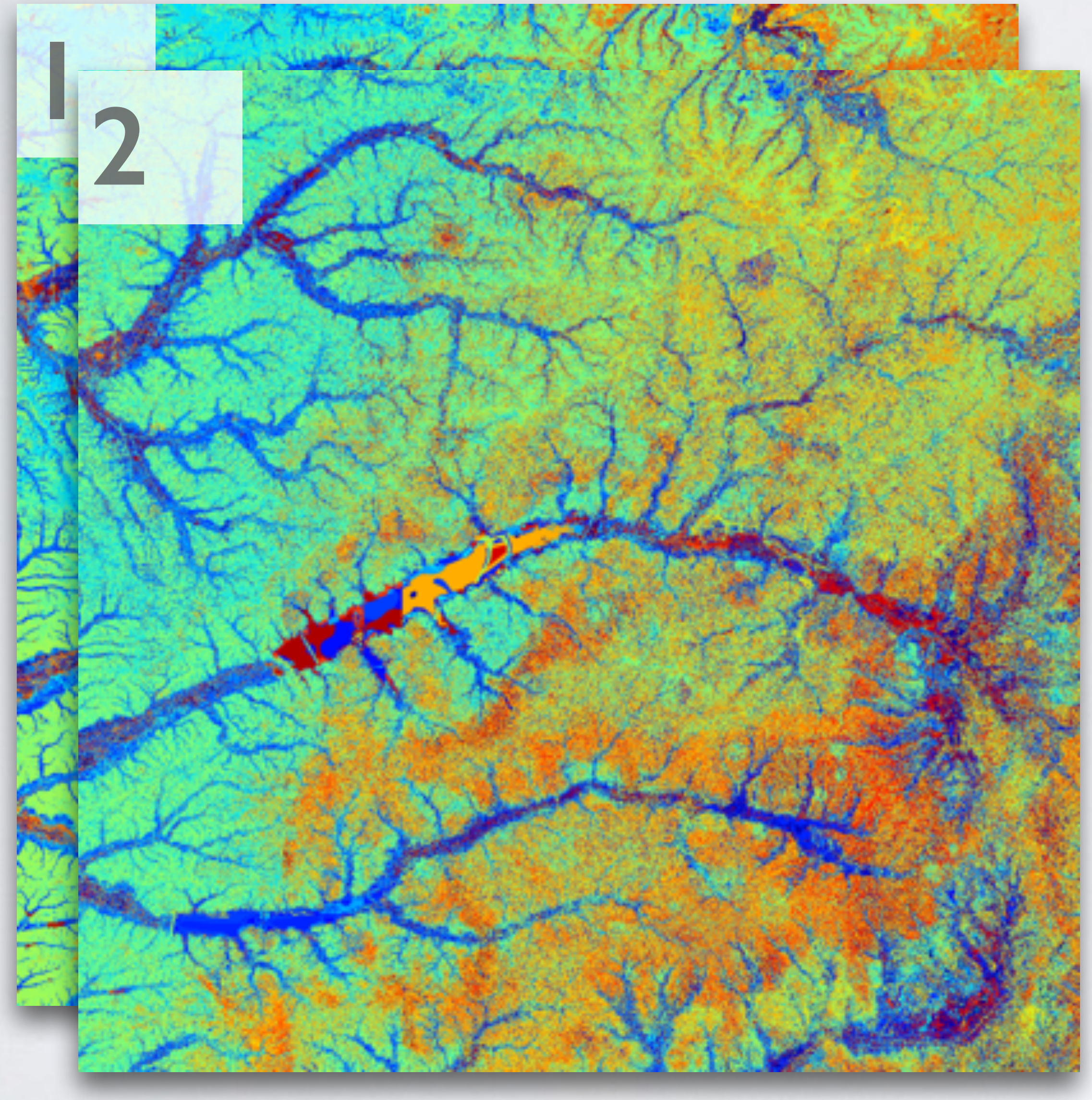
PROBABILITY RANKED

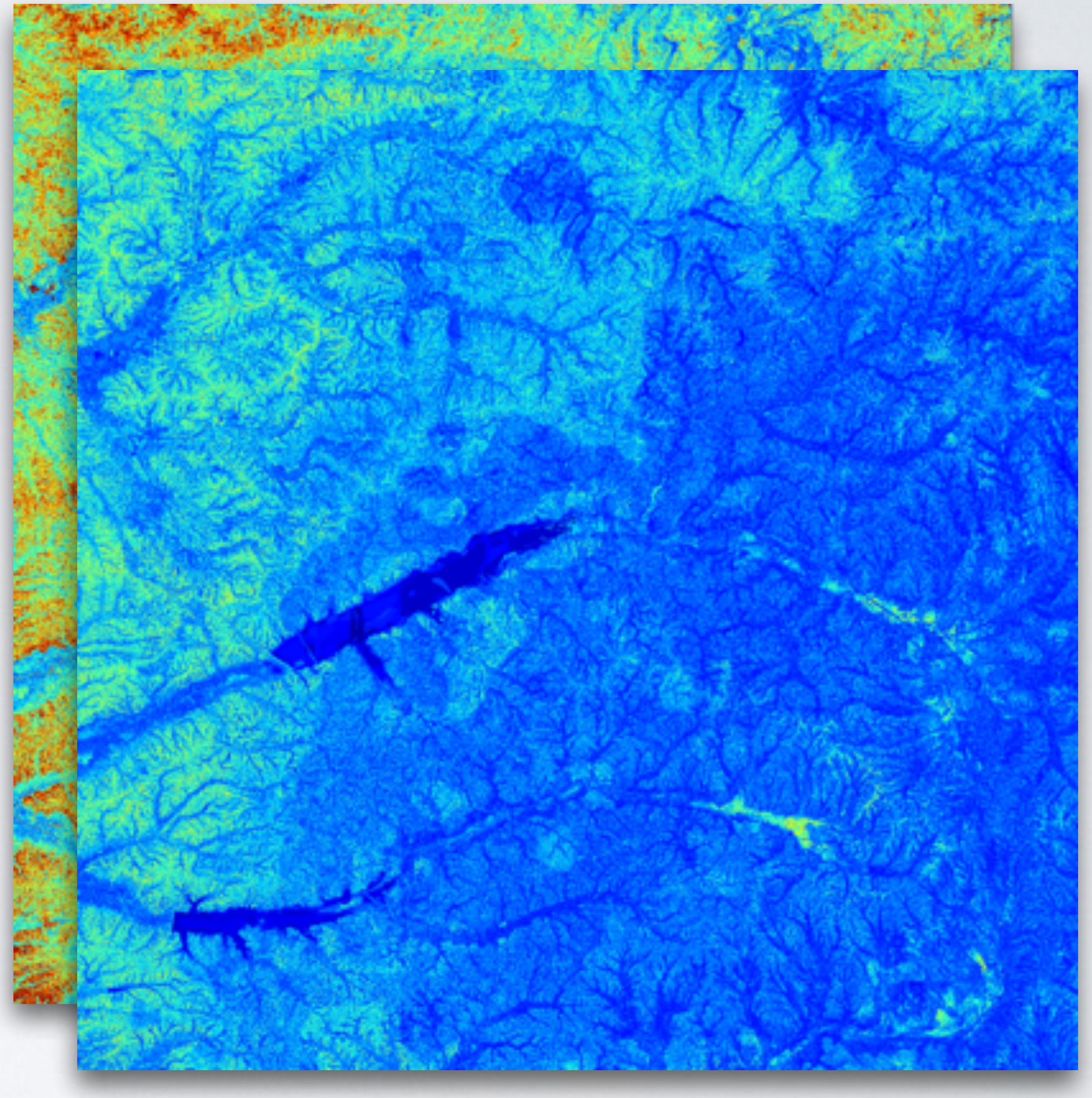
Component ← Probability

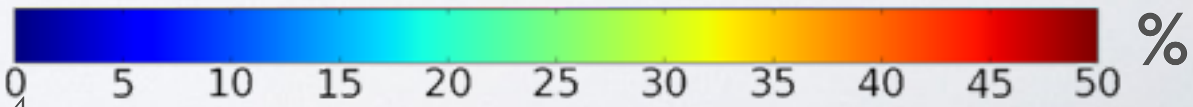


PROBABILITY RANKED

Component ← Probability

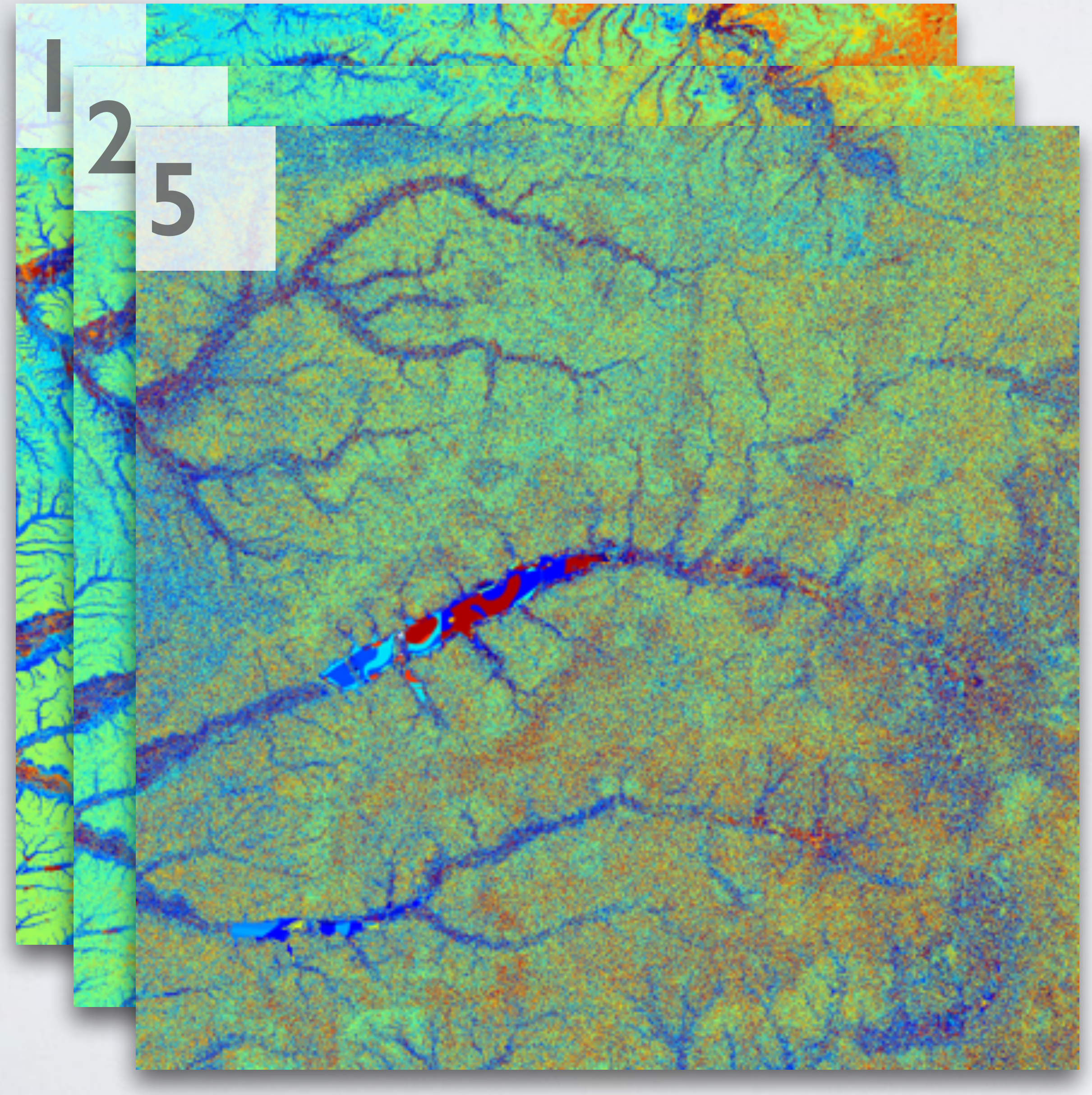


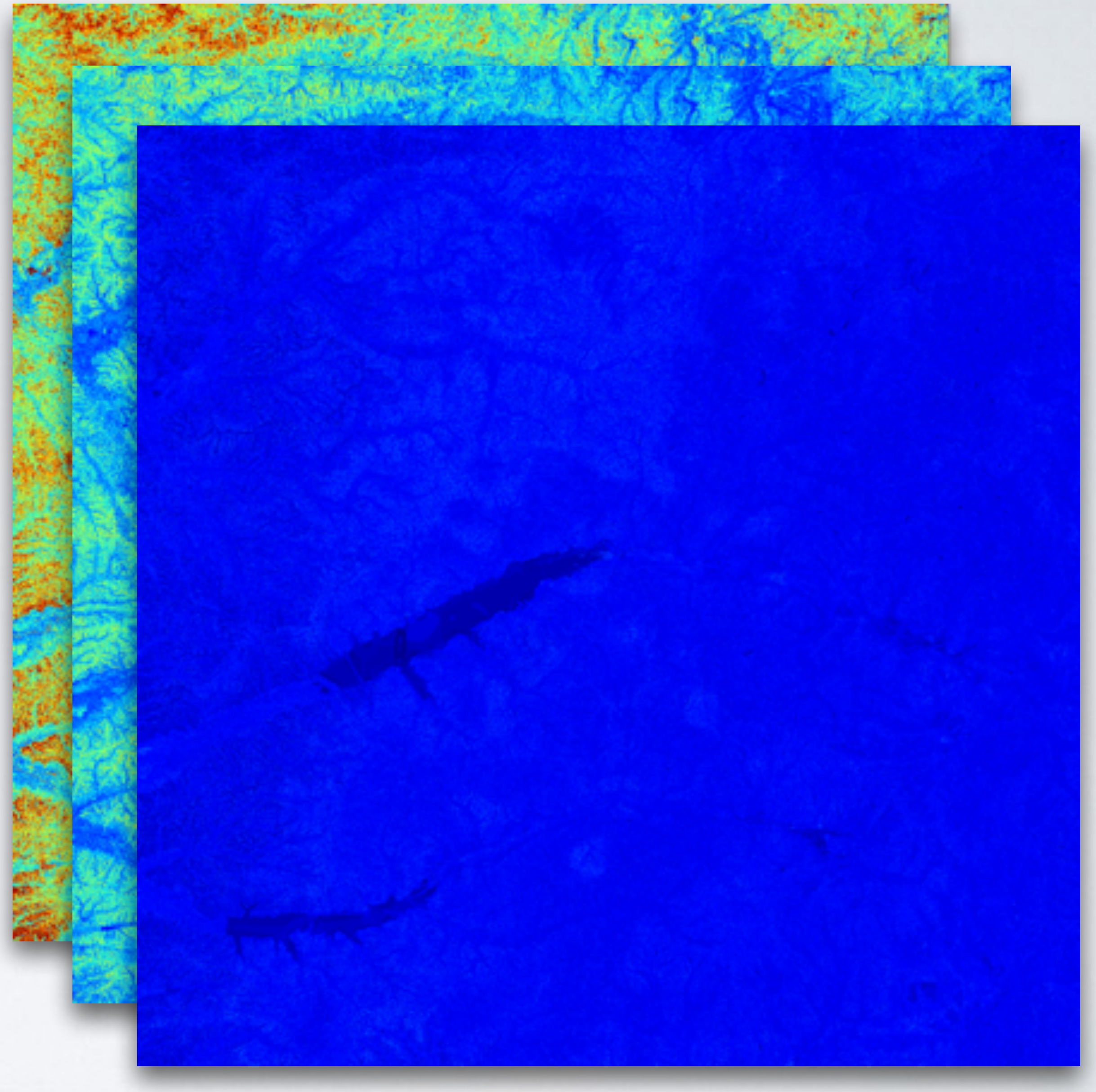


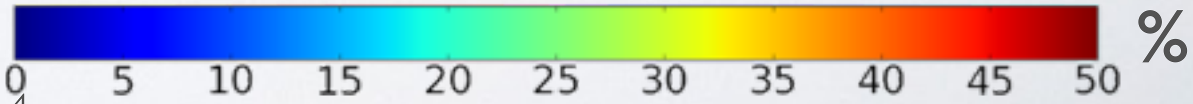


PROBABILITY RANKED

Component ← Probability

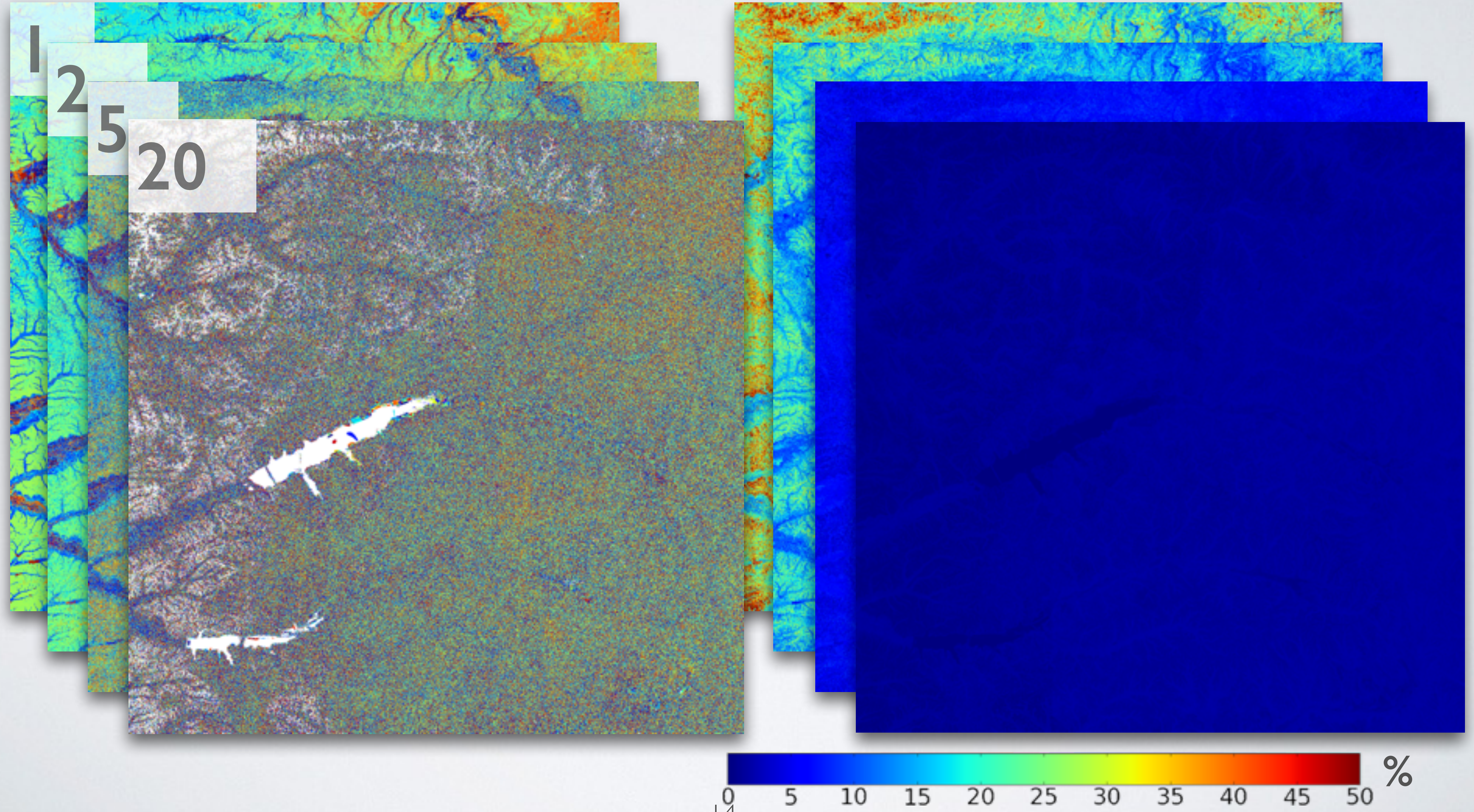






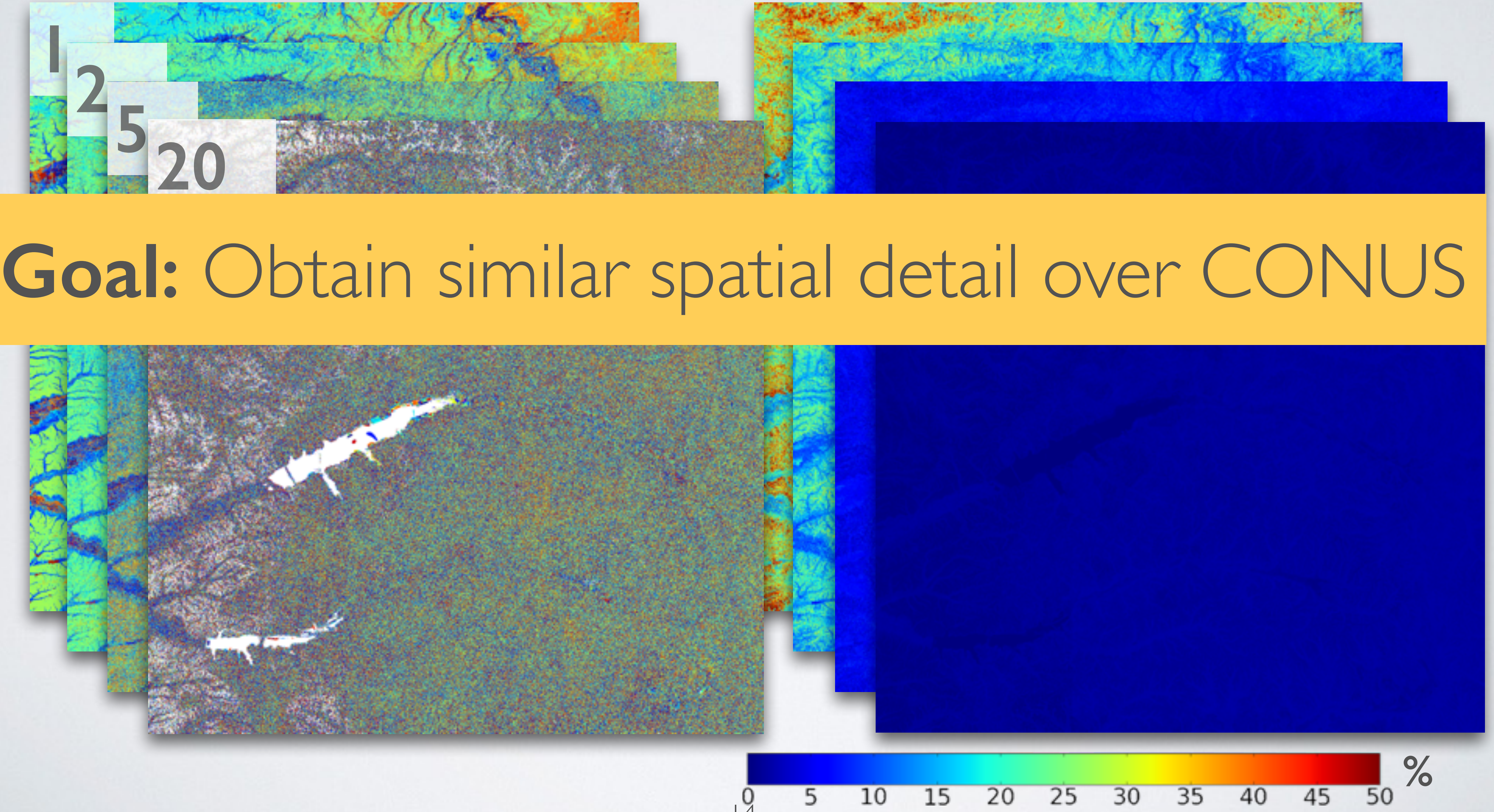
PROBABILITY RANKED

Component ← Probability



PROBABILITY RANKED

Component ← Probability



Goal: Obtain similar spatial detail over CONUS

Application over CONUS

CONUS 30 meters → ~9 billion grid cells

Feasible Approach: Moving window

- Split up domain into overlapping blocks
- Run DSMART on each block
- Small region → small sample size → fast random forest
- ~25,000 blocks → **500,000 core hours**

High Performance Computing: Blue Waters

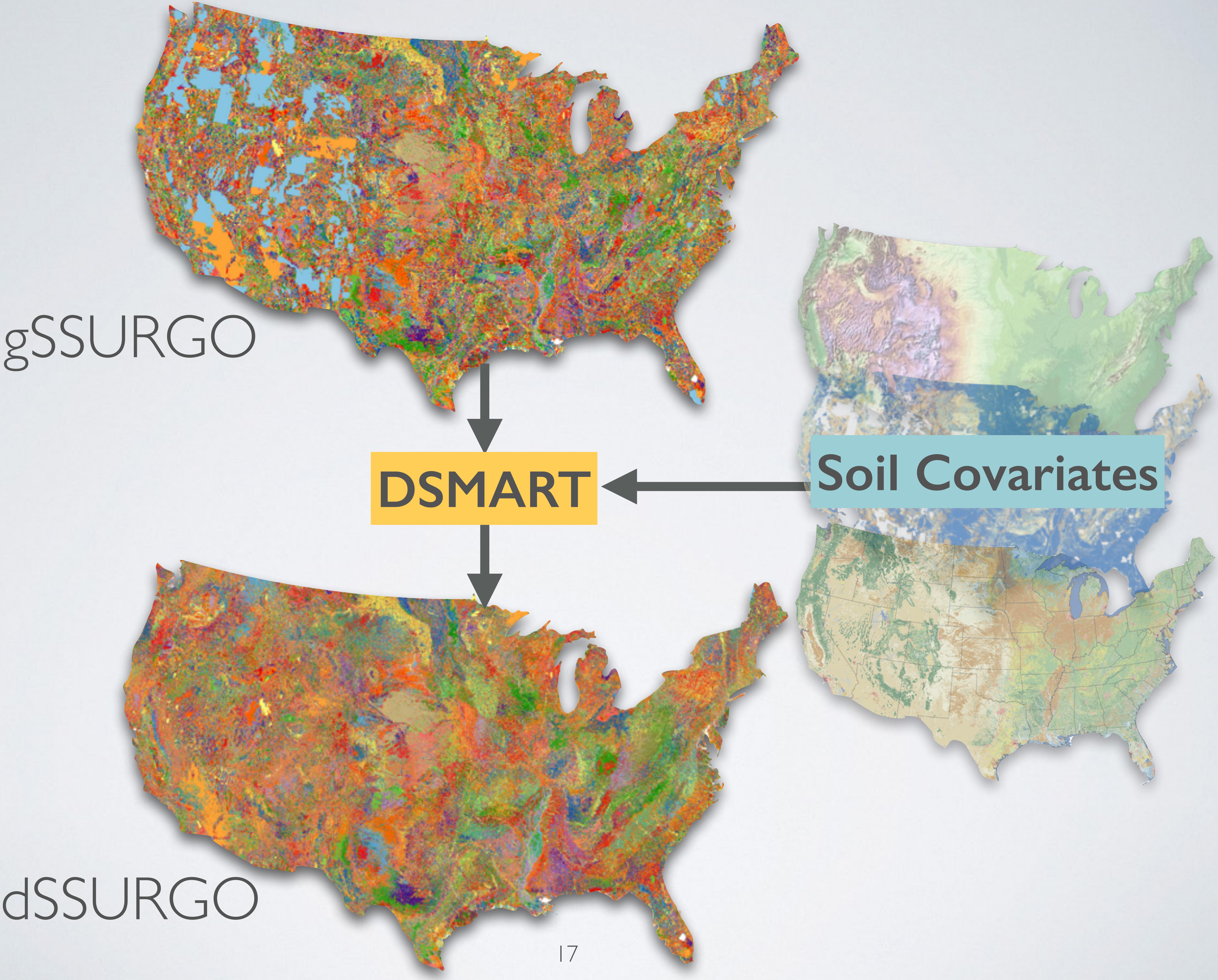
	Machine Stats	Comparison
Number of Cores	600,000	>13 quadrillion calculations per second
Memory	1.5 petabytes	300 million images
Short Term Storage	25 petabytes	All printed documents in all libraries
Long Term Storage	500 petabytes	10% of all words spoken by humankind

Source: NCSA

500,000 hours (57 years)

↓

5 hours



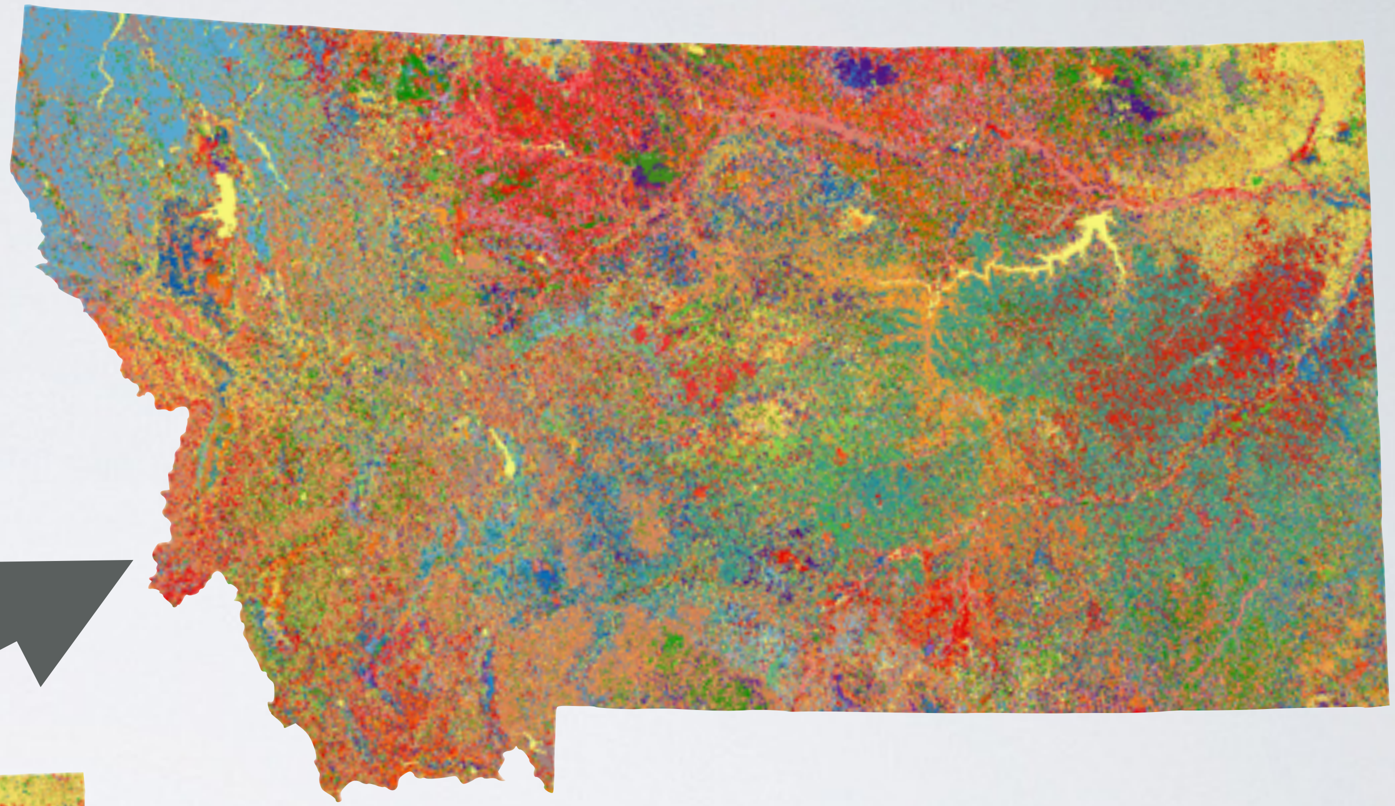
gSSURGO

DSMART

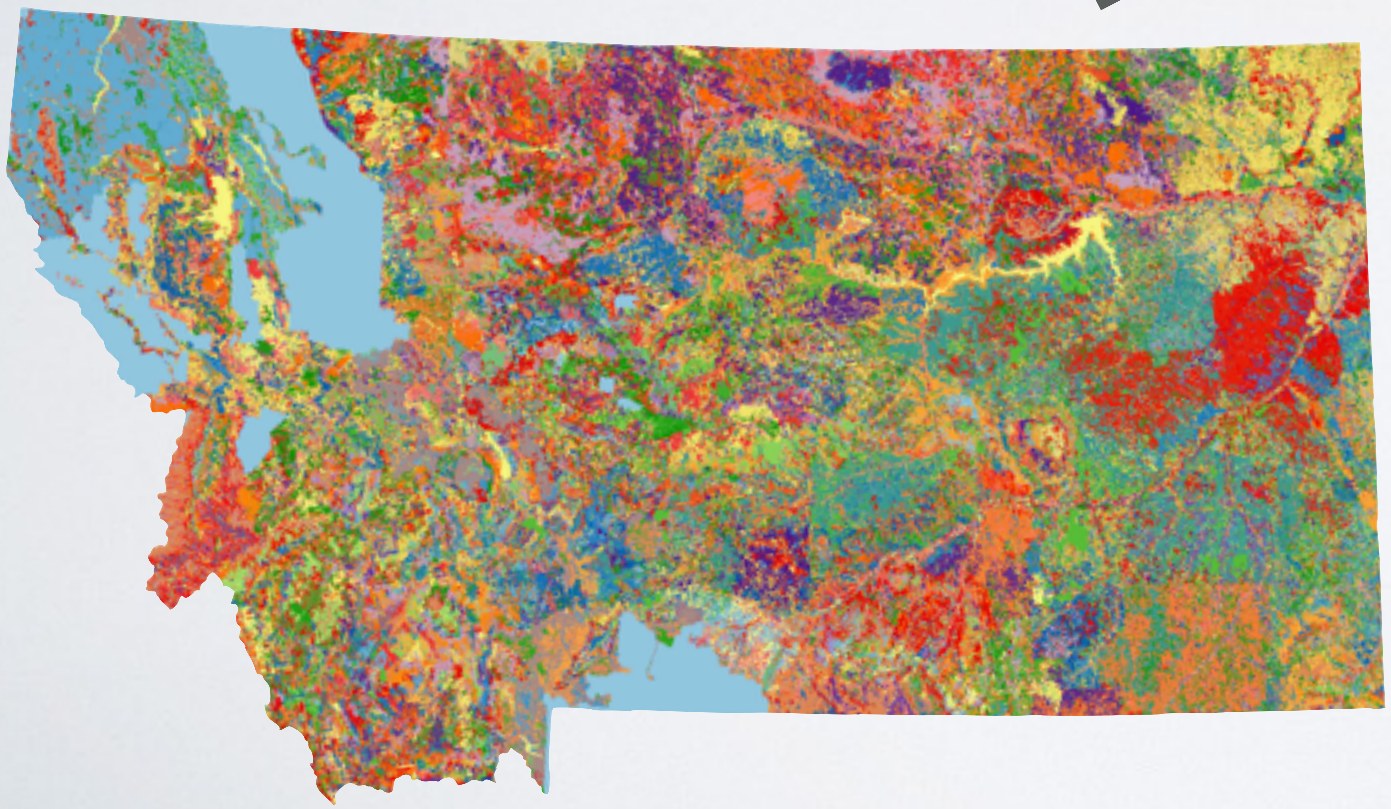
Soil Covariates

dSSURGO

DSMART: Montana

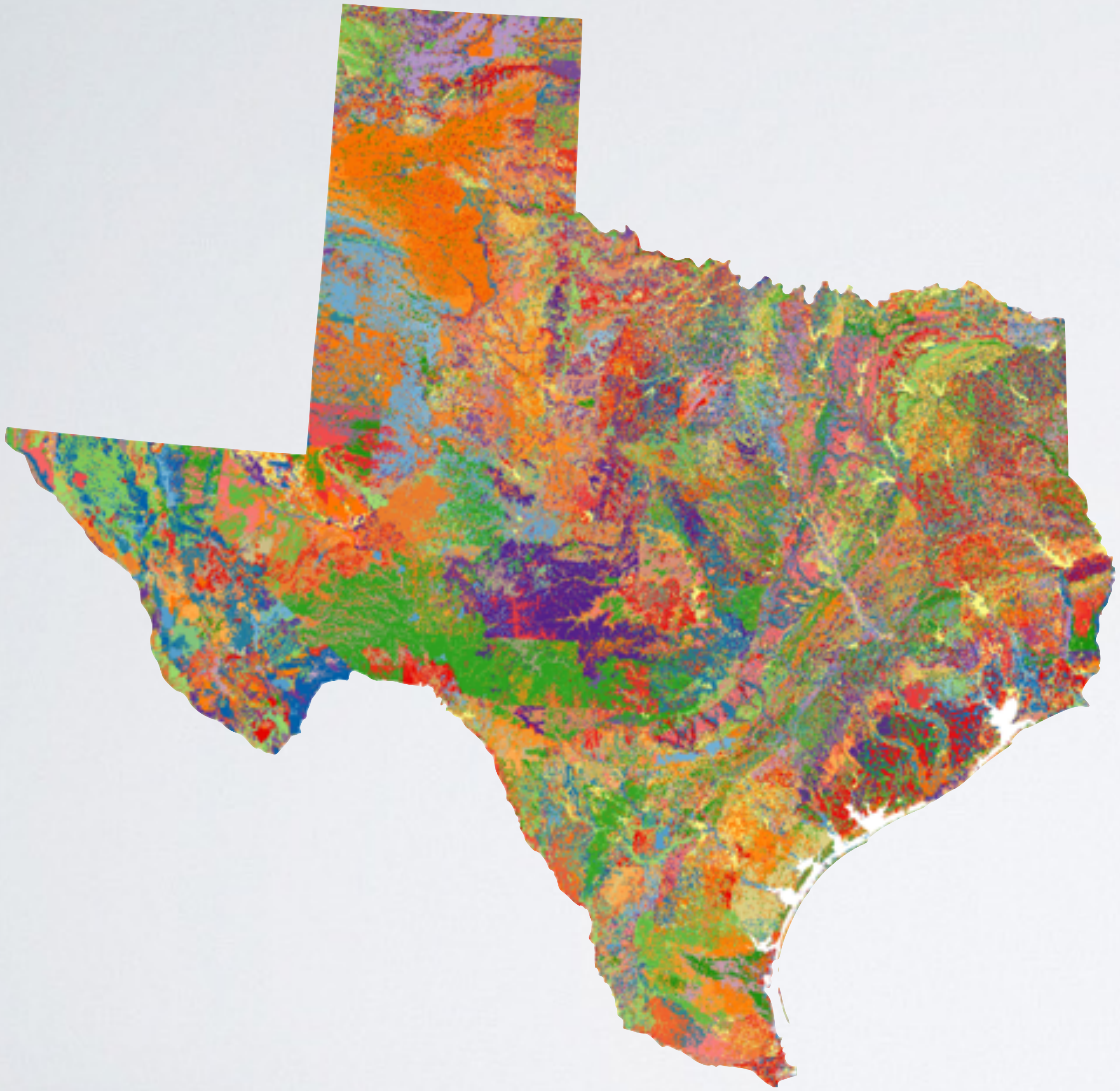


dSSURGO

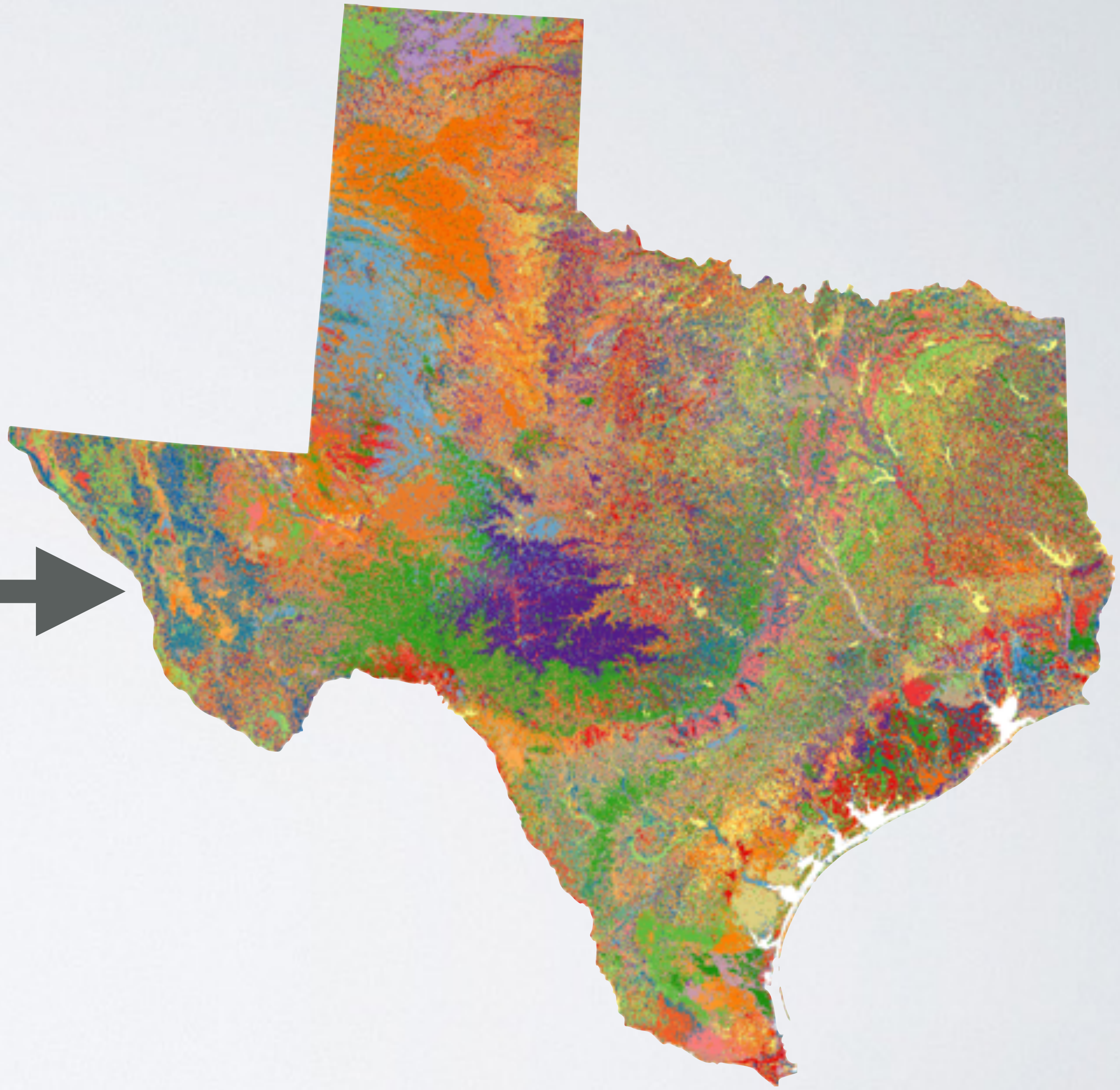


gSSURGO

DSMART: Texas

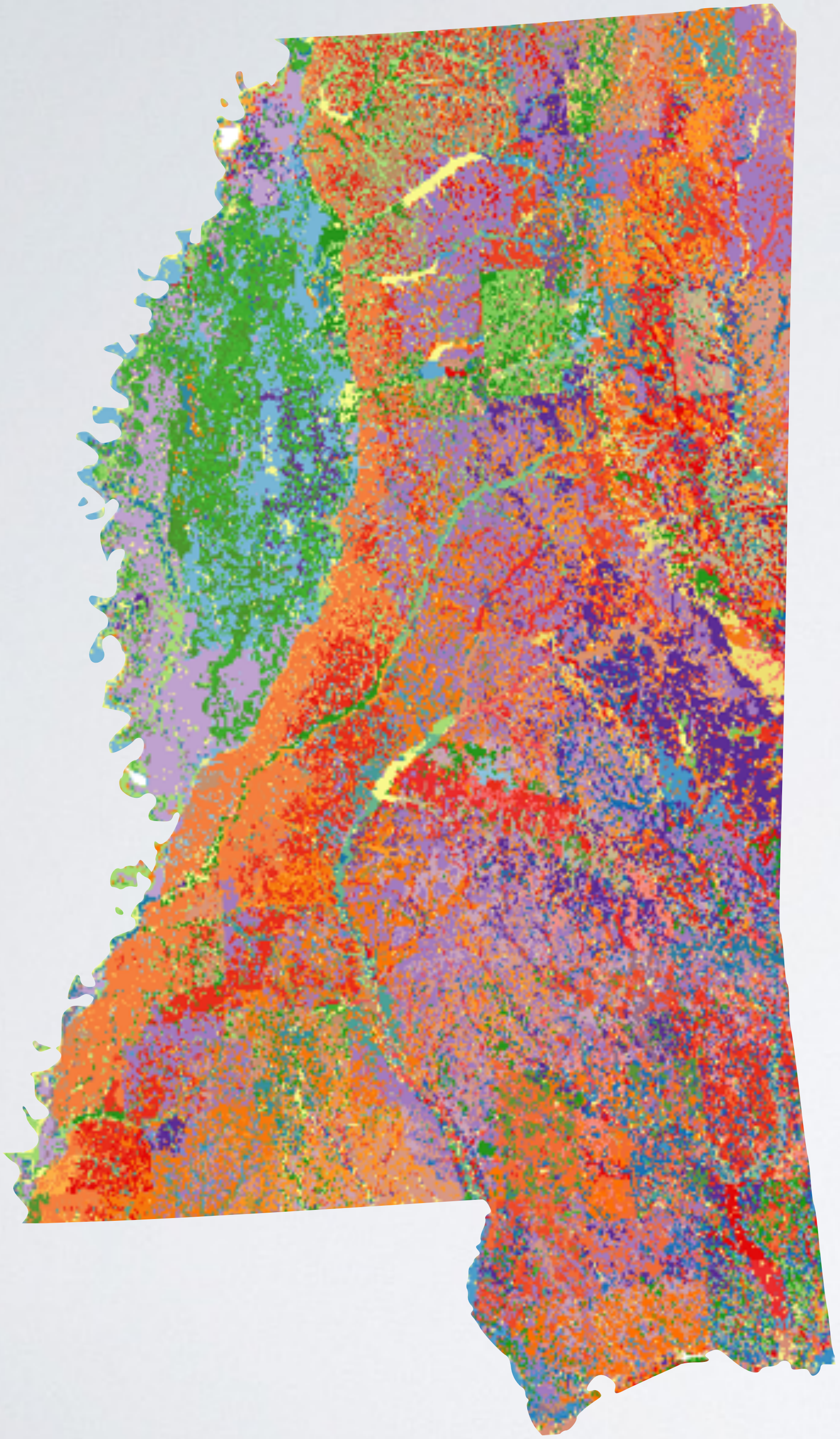


gSSURGO

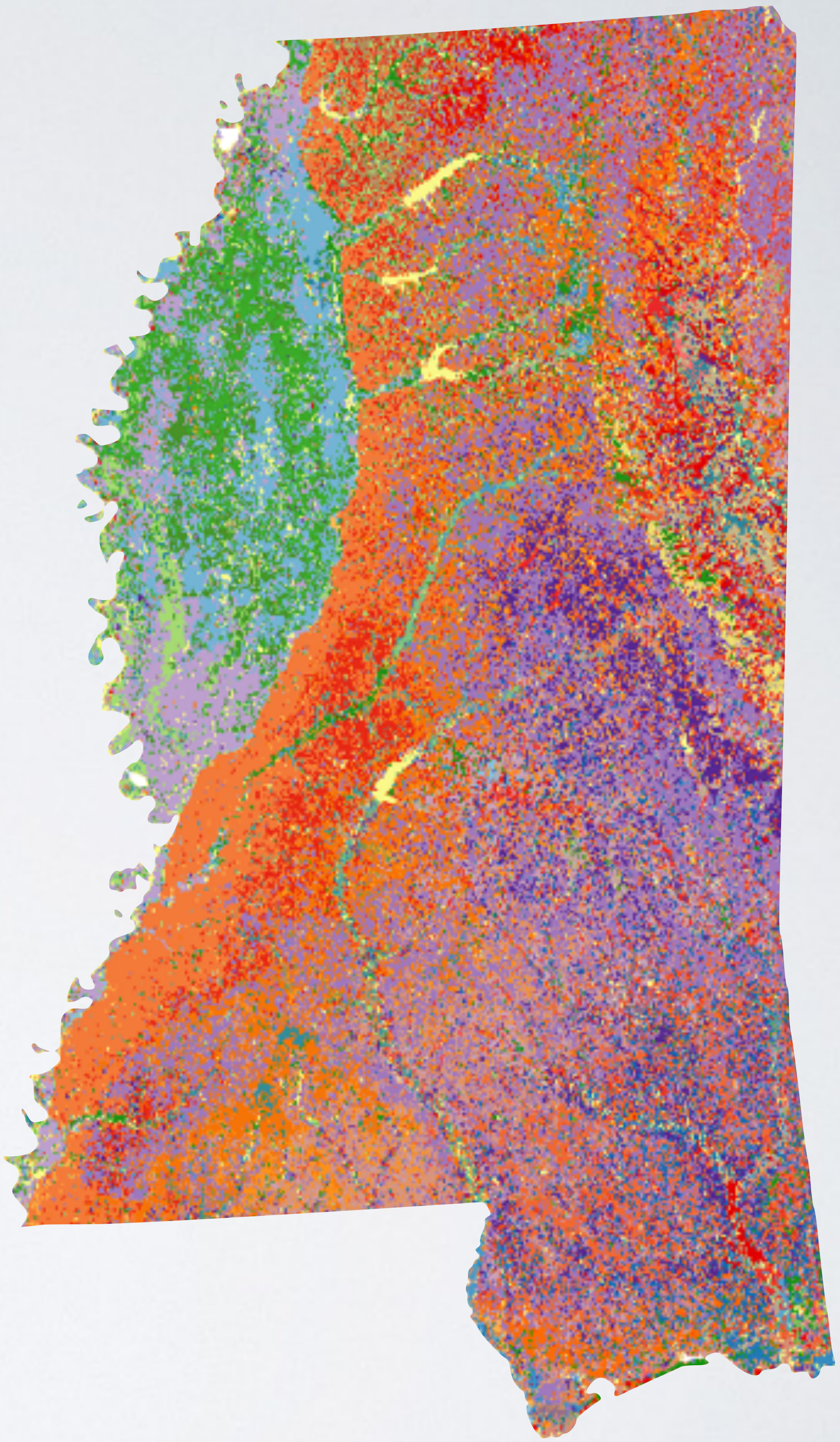


dSSURGO

DSMART: Mississippi

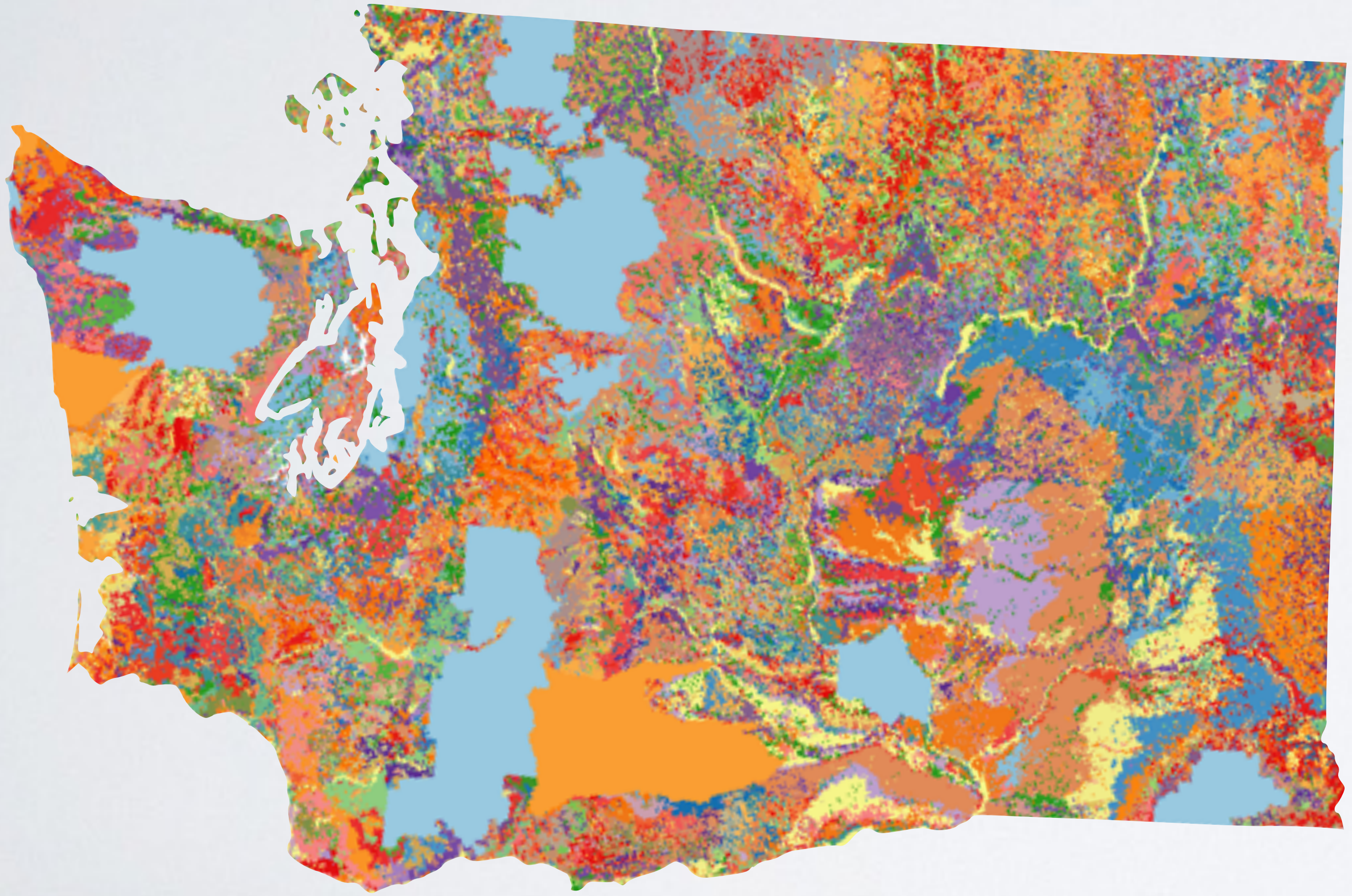


gSSURGO

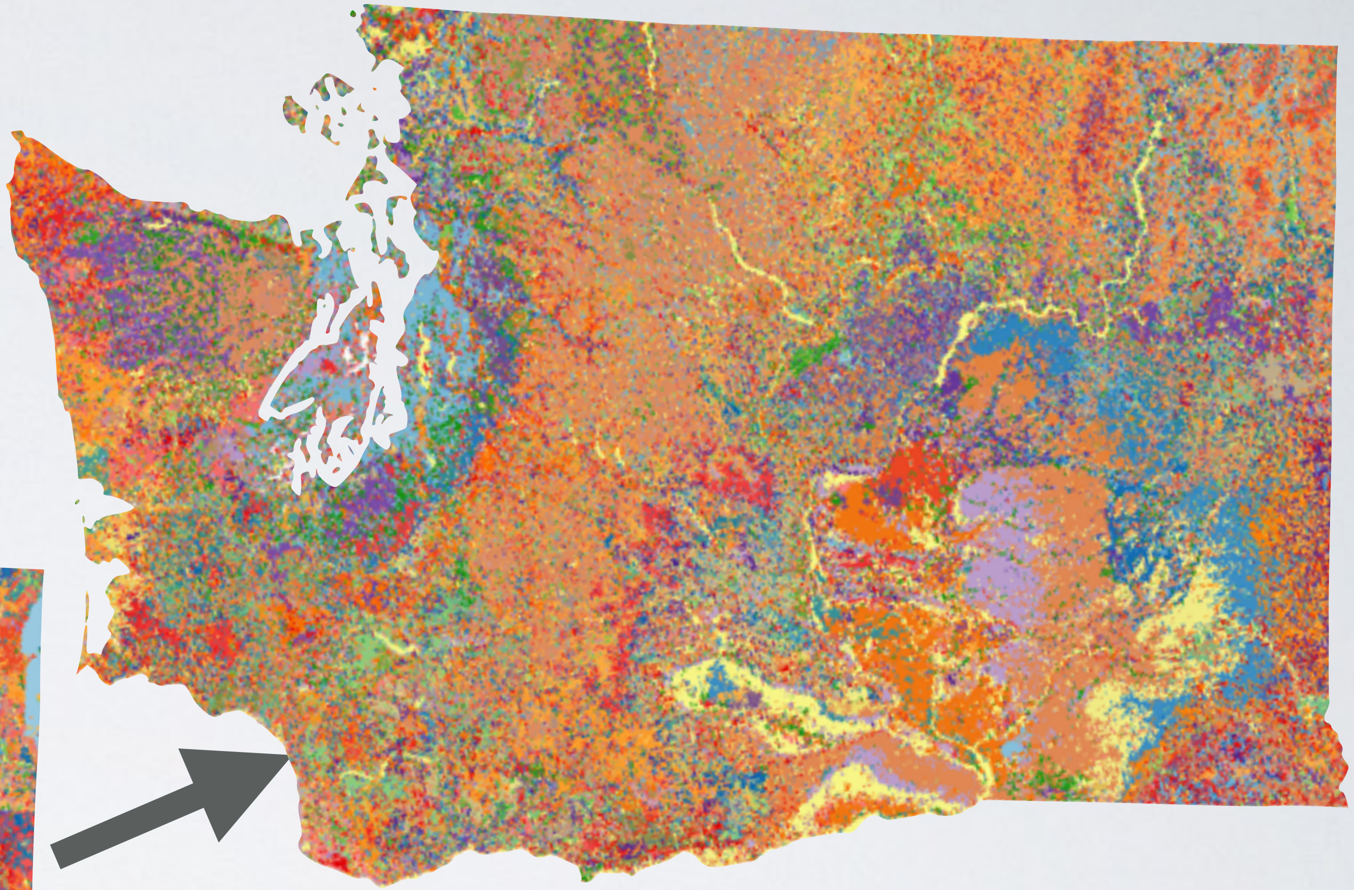


dSSURGO

DSMART: Washington

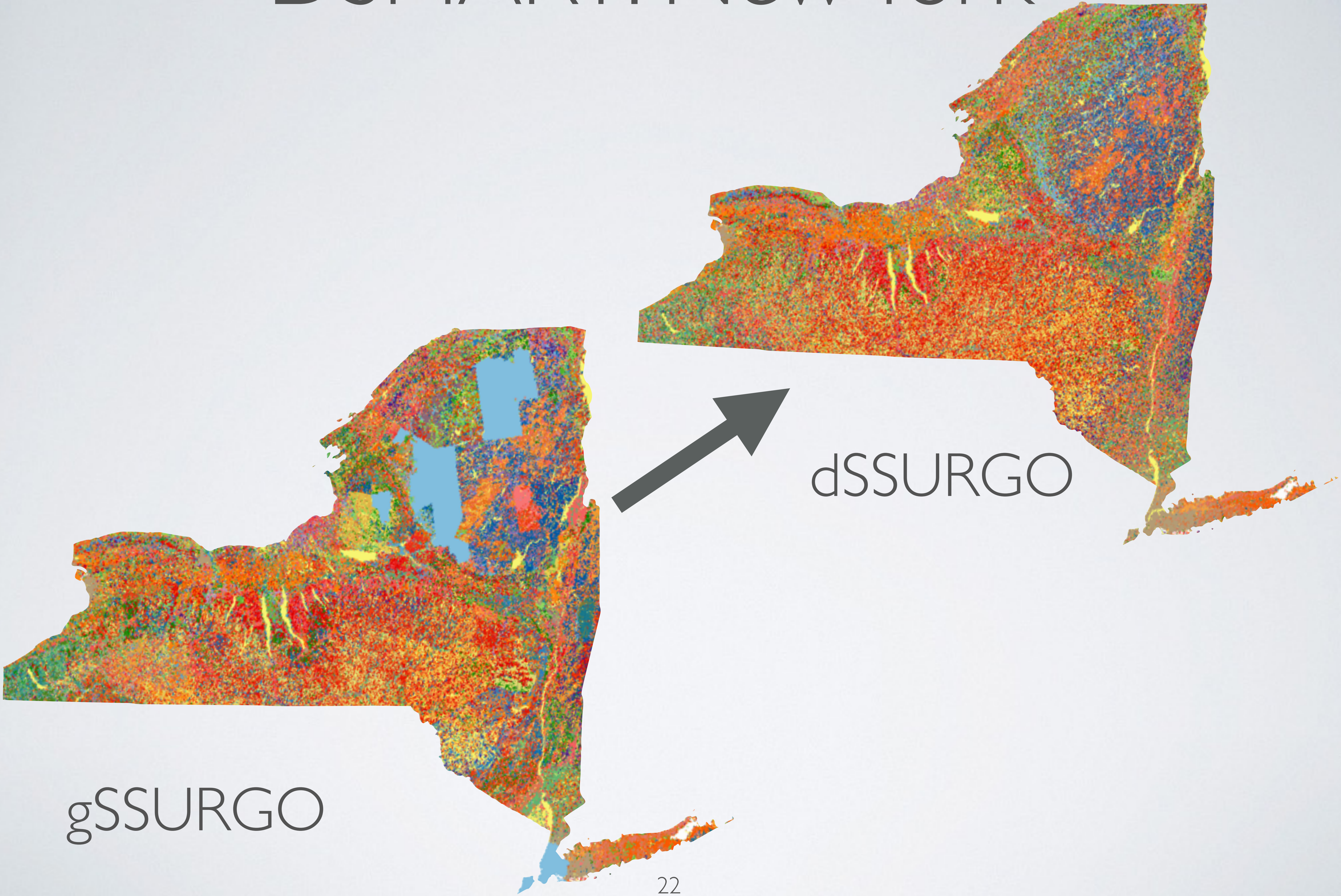


gSSURGO



dSSURGO

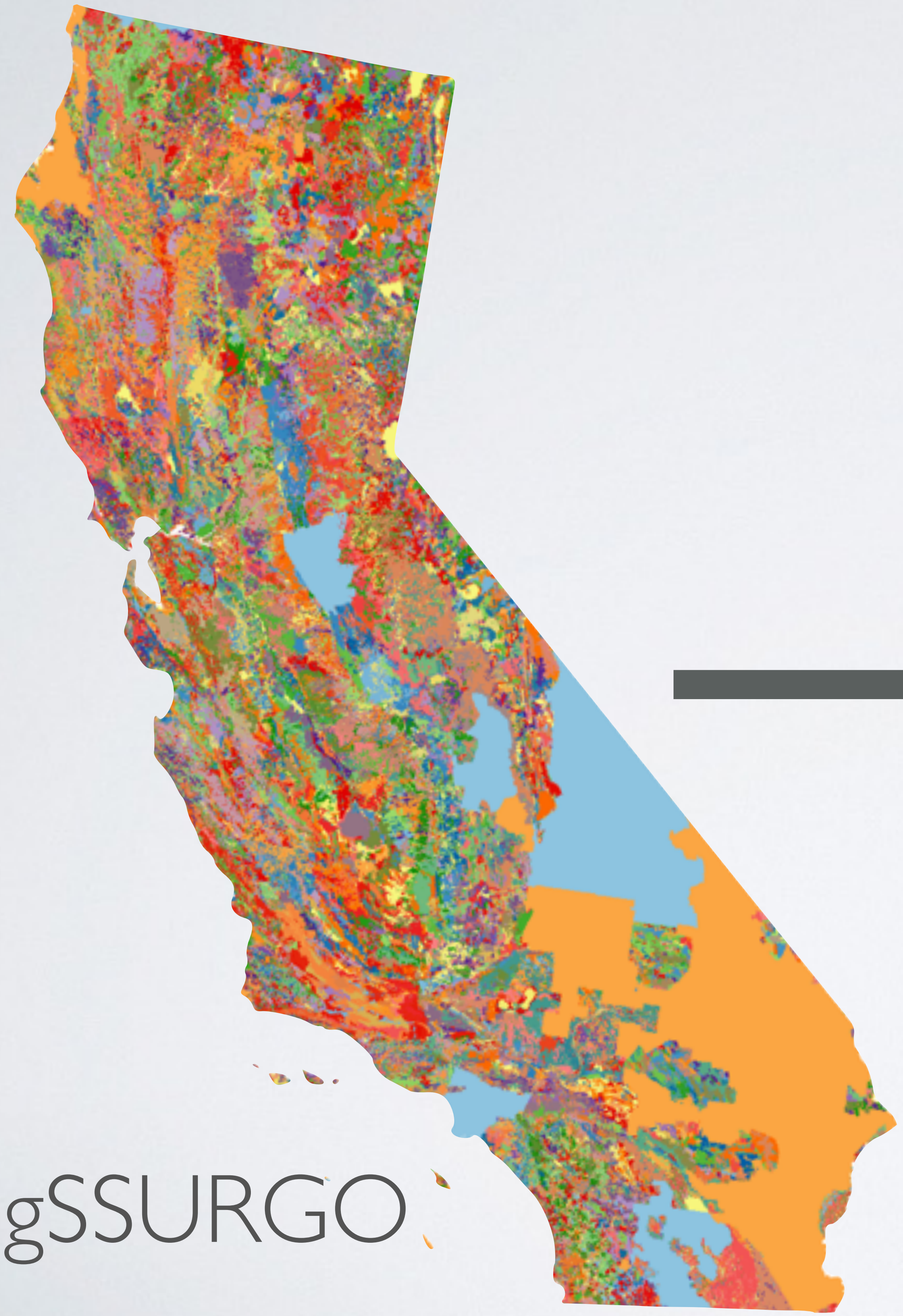
DSMART: New York



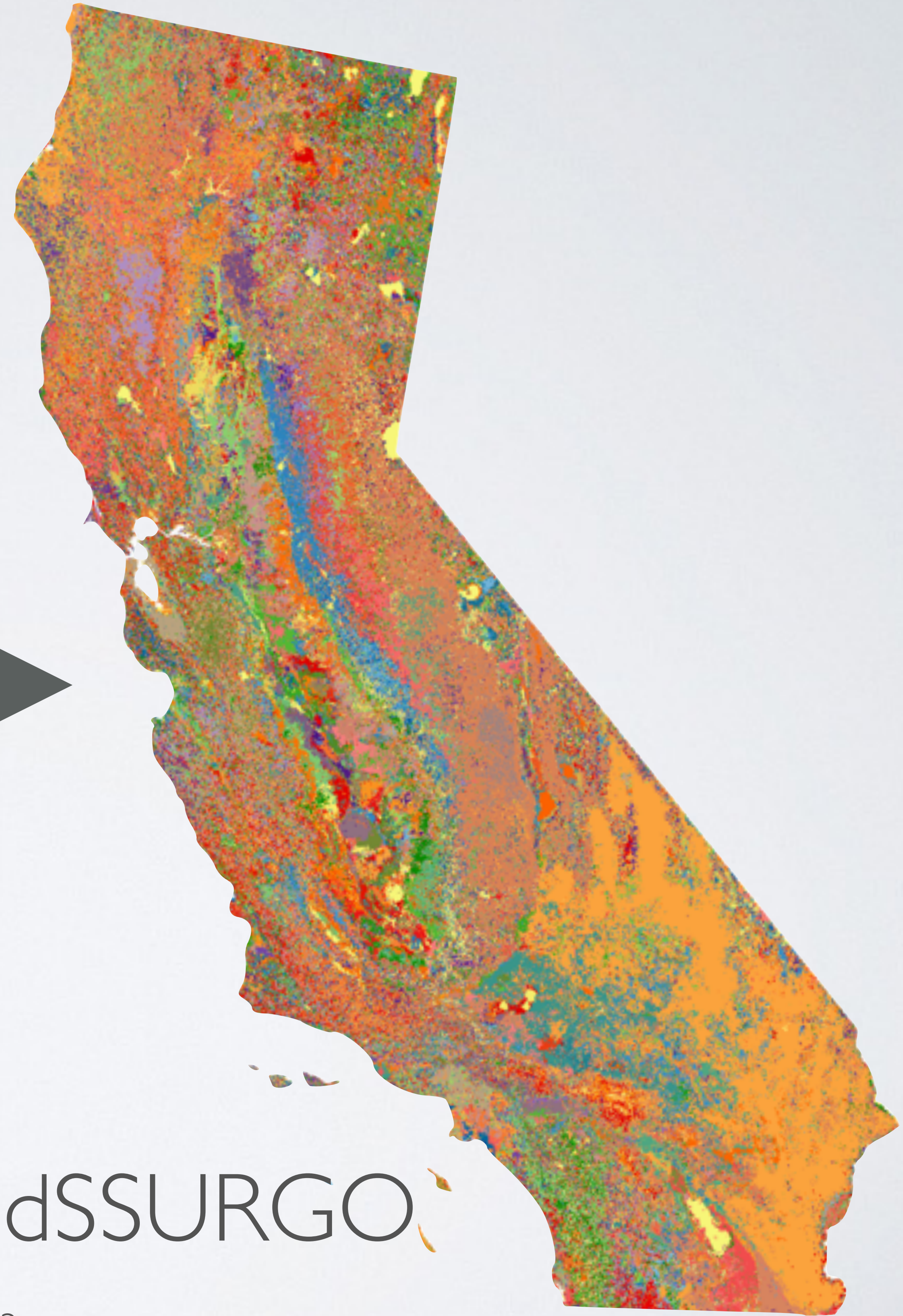
dSSURGO

gSSURGO

DSMART: California

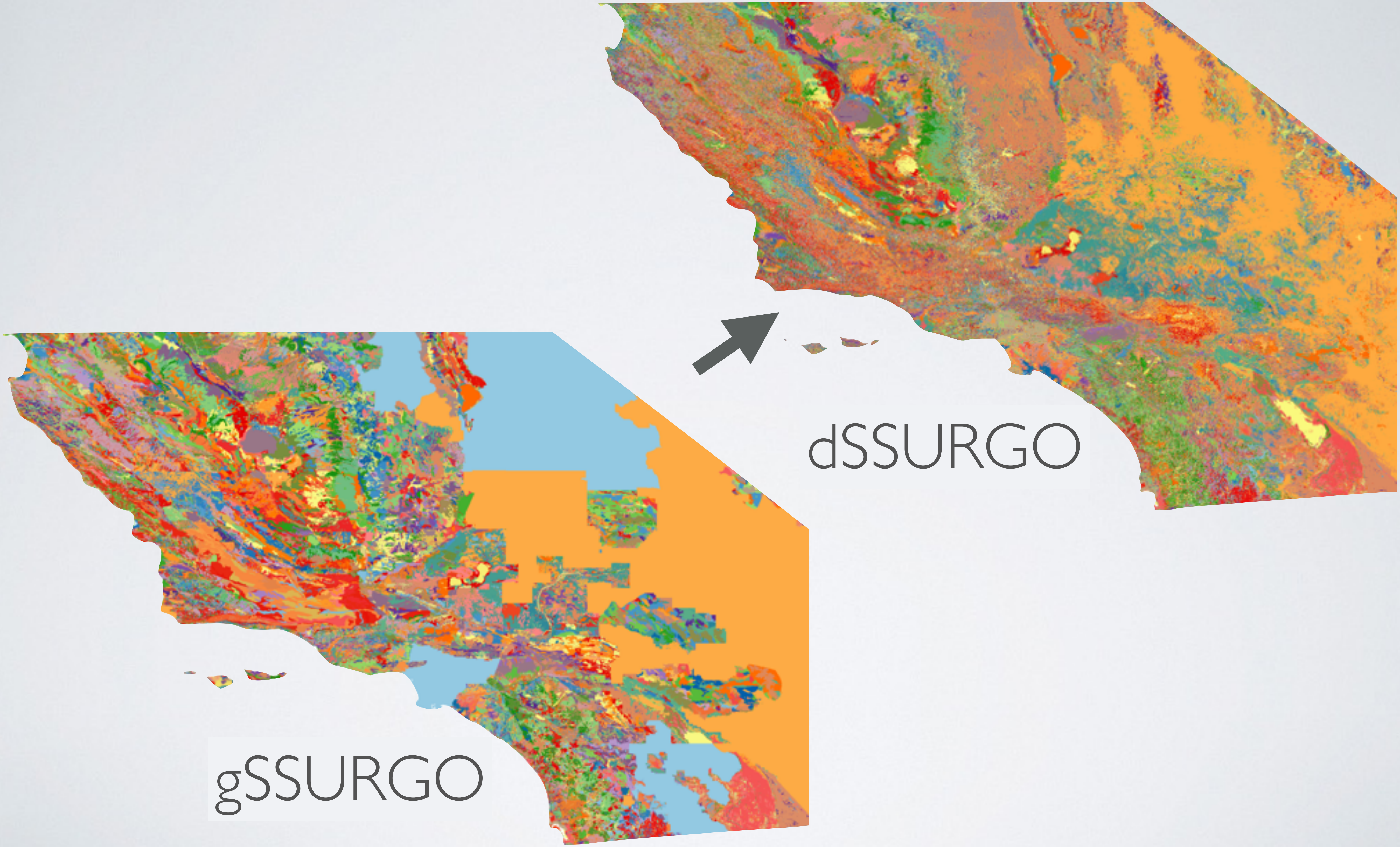


gSSURGO

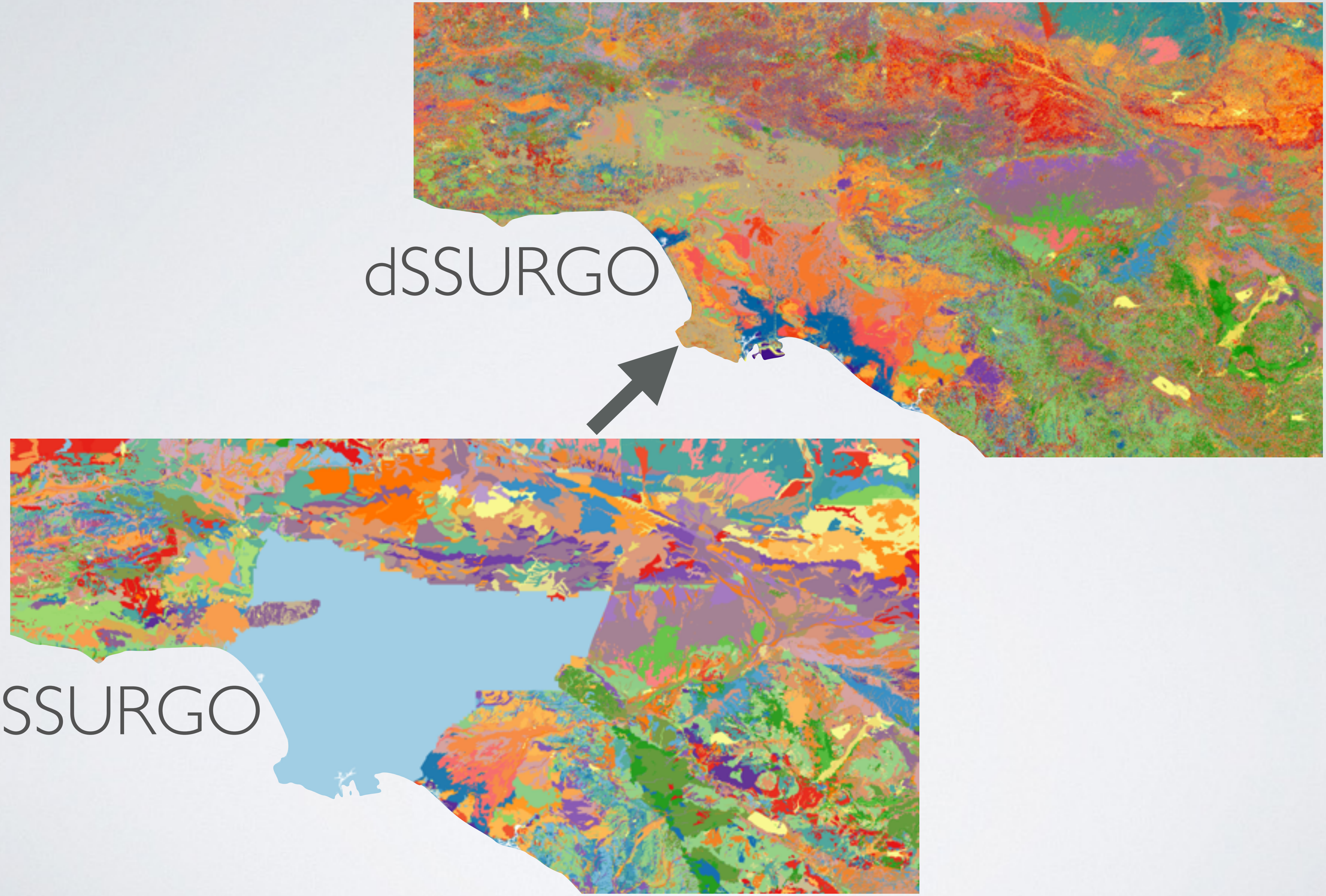


dSSURGO

DSMART: Southern California



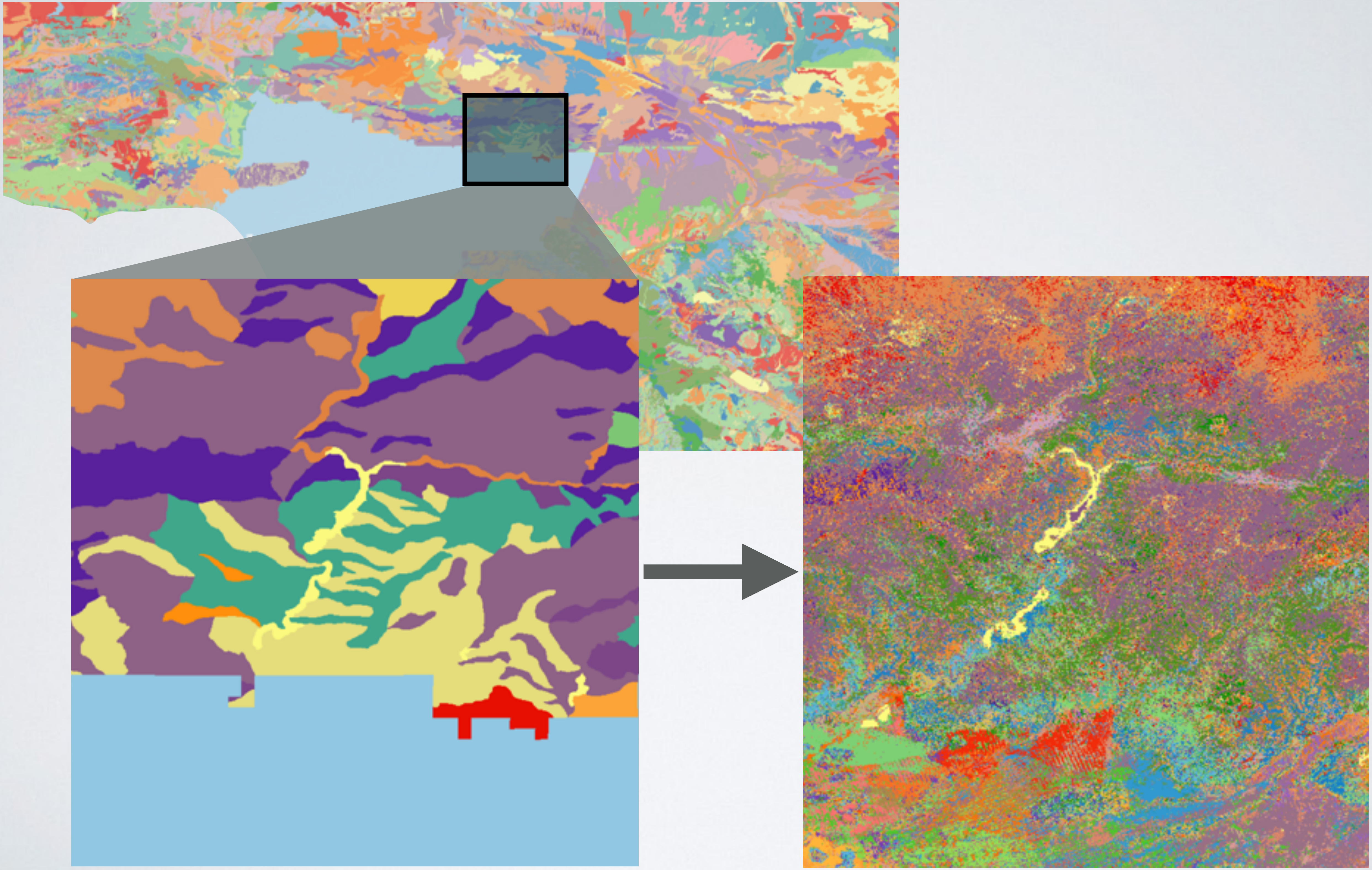
DSMART: Greater LA Area



dSSURGO

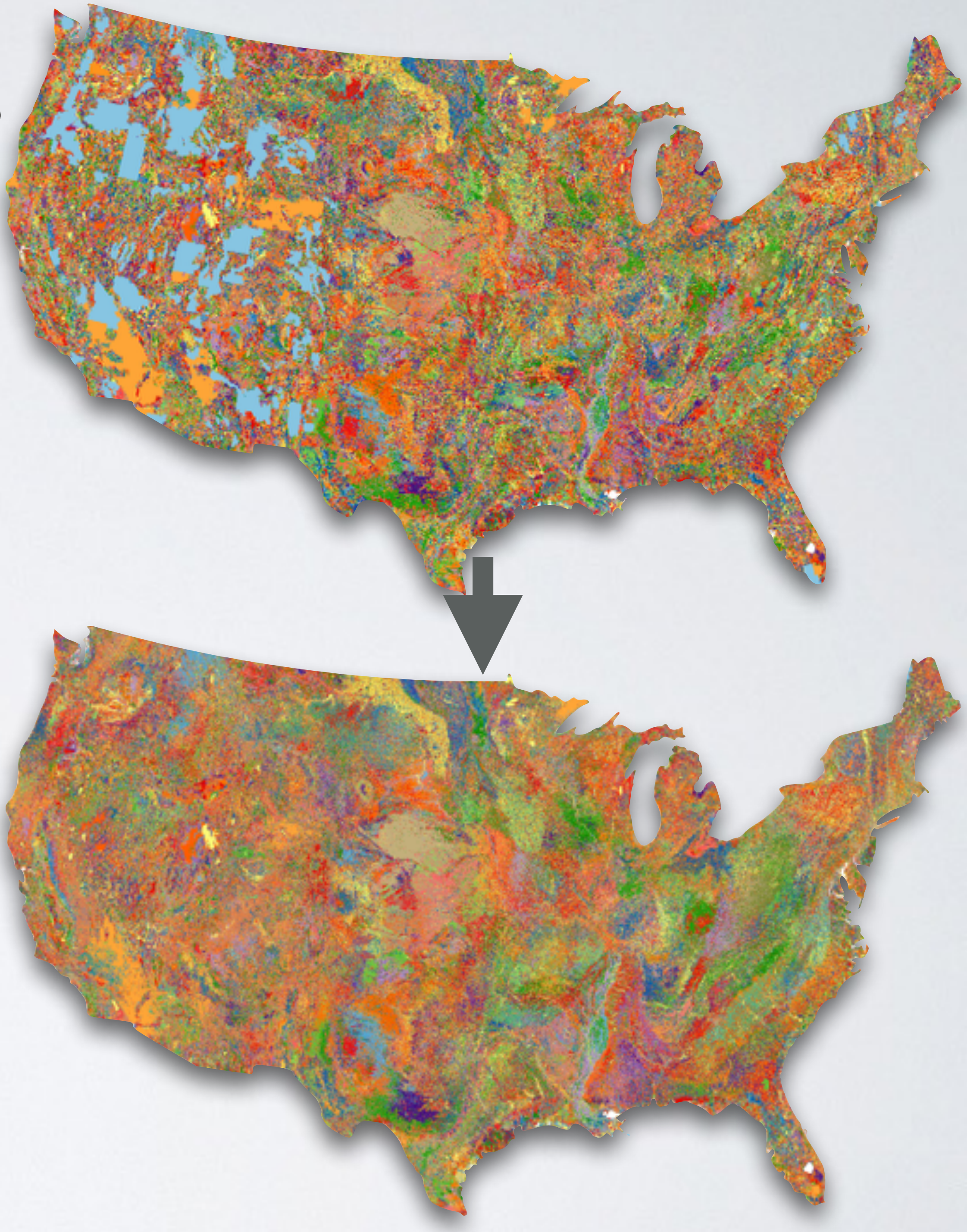
gSSURGO

Angeles National Forest



Conclusions and Next Steps

- **dSSURGO** - CONUS at 30 meters
  - 50 most probable components (and probabilities)
  - ~2 terabyte dataset (freely accesible)
  - [stream.princeton.edu/dSSURGO]()
- **Next Steps**
  - Applications (e.g. Hydrologic Modeling)
  - Validation (Need your help!)



PRINCETON UNIVERSITY NRCS THE UNIVERSITY OF SYDNEY BLUE WATERS

Conclusions and Next Steps

- **dSSURGO** - CONUS at 30 meters
  - 50 most probable components (and probabilities)
  - ~2 terabyte dataset (freely accesible)
  - stream.princeton.edu/dSSURGO
- **Next Steps**
  - Applications (e.g. Hydr
  - Validation (Need your help!)

Questions?

PRINCETON UNIVERSITY NRCS THE UNIVERSITY OF SYDNEY BLUE WATERS